UNIVERSITÄTSMEDIZIN GÖTTINGEN ALLGEMEINMEDIZIN UMG

Können ausgewählte ambulante Routinedaten anonym sein?

Hauswaldt J, und die Konsortionalpartner im RADARplus Projekt
Universitätsmedizin Göttingen, Institut für Allgemeinmedizin

Hintergrund

Anonyme Datennutzung ist nur zulässig, wenn die Daten eines Datensatzes nicht oder nur mit unverhältnismäßigem Aufwand, auch nicht nachträglich, auch nicht durch Verknüpfung mit weiteren Daten oder Informationen, auf eine Person bezogen werden können (DSGVO, Erwägungsgründe 26).

- Können Routinedaten aus Arztpraxisinformationssystemen "faktisch anonymisiert" werden, um sie sekundär für Versorgungsforschung zu nutzen?

RADARplus
Anonymisierte Routinedaten aus der ambulanten Versorgung für die Versorgungsforschung

Methode

Drei vorliegende Datensammlungen hausärztlicher Routinedaten werden erneut untersucht. Das Geburtsdatum einer Person ist jedenfalls als quasi-identifizierend anzusehen, deshalb wird es auf "Geburtsjahr" trunkiert. ICD-Diagnosen werden auf drei führende Stellen verkürzt und je Person quartalsweise gezählt.

- Zunächst werden "**Geschlecht**" und "**Geburtsjahr**" unter den Datenschutzmodellen "k-Anonymität" und „Durchschnittliches Re-Identifizierungsrisiko" untersucht, und zwar in der Ausgangsdatensammlung sowie in einer sinnvollen und wünschenswerten Modifikation nach Transformation mittels der Open-Source Software "ARX – Data Anonymization Tool", Version 3.8.0.
- Ergebnisse für verschiedene Risiko-Aspekte der Re-identifizierung sowie für den Informationsverlust und für die Brauchbarkeit weiterer Datenschutzmodelle werden einander gegenübergestellt.
- Anschließend werden die drei Datensammlungen unter Hinzunahme der 6 häufigsten "**ICD-Dreisteller**“ je Patient in gleicher Weise erneut untersucht.

ARX

Ergebnisse

Daten-sammlung	Anzahl Patienten	Anzahl Praxen	Zeitraum	Datenschutzmodelle				Generalisierung *Clustering, microaggregation*	Optimale Transformation	Datenqualität *Granularity*
1	426.374	152	1994 - 2007	k-Anonymität	k = 30	Durchschnittliches Re-Identifikations-risiko	(0,001)	Geschlecht: Original | *	[0,1,1,0,0,0,0,0]	72,2 %
2	3.704	1	1994 - 2017	k-Anonymität	k = 30	Durchschnittliches Re-Identifikations-risiko	(0,1)	Geburtsjahr: Original | 5 | 10 | 20 | *	[0,3,1,0,0,0,0,0]	61,3 %
3	91	7	2012 - 2019	k-Anonymität	k = 5	Durchschnittliches Re-Identifikations-risiko	(0,1)	ICD3: A00 | A** | ***	[0,2,1,2,2,2,2,2]	28,2 %

Abb. 1 Drei Datensammlungen

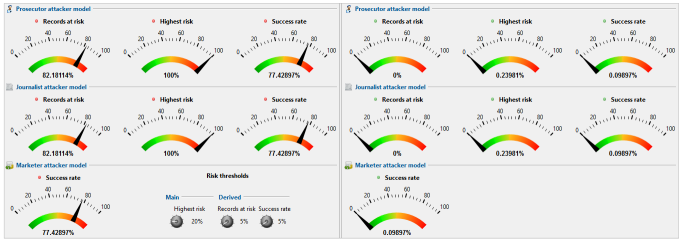

Abb. 2 Datensammlung 1, original und transformiert, drei Angriffsszenarien

- Größere Datensammlungen besser noch als kleine können bereits mit wenig eingreifenden, gezielten Modifikationen und entsprechend geringem Informationsverlust wirksam anonymisiert werden.
- Eine absolute Anonymisierung gelingt jedoch nicht.

Diskussion und Schlussfolgerungen

Eine erfolgreiche Anonymisierung, wenn überhaupt sinnvoll möglich, kann keinesfalls "absolut", also dichotom ("anonym ja oder nein") gelingen. Sie ist vielmehr in Abstufungen im Spannungsfeld zwischen (akzeptiertem) Risiko der Re-Identifizierung und dem (noch sinnvollen) Informationsverlust zu bewerten, bevor eine Sekundärnutzung erfolgt.

Ob mit ausreichender Wahrscheinlichkeit ein Personenbezug nicht hergestellt werden kann („faktische Anonymisierung“), muss immer im konkreten Einzelfall der untersuchten abgeschlossenen Datensammlung und unter Berücksichtigung der technischen und organisatorischen Maßnahmen ihrer Gewinnung und Verarbeitung (Forschungsinfrastruktur) sowie möglicher Angriffsszenarien nachgewiesen und beurteilt werden.