UNIVERSITÄTSMEDIZIN GÖTTINGEN ALLGEMEINMEDIZIN UMG

Zur De-Identifizierung von Feldinhalten in hausärztlichen Routinedaten

Exemplarische Untersuchung an BDT-Daten einer Hausarztpraxis, 1994 bis 2017

Hauswaldt J, Groh R, Kaulke K, Schlegelmilch F, Zarei A, Hummers E
für AGENS Methodenworkshop, 24. und 25. 02.2022, Frankfurt / Main

johannes.hauswaldt@med.uni-goettingen.de

Hintergrund und Ziel

- Sekundäre Nutzung von hausärztlichen Routinedaten ist technisch und organisatorisch rechtskonform machbar
- Anonymes Szenario bei bestehender Rechtslage nicht erlaubt
- Potentiell identifizierende Feldinhalte (PIF) behindern die „faktische Anonymisierung“ eines wissenschaftlich genutzten Sekundärdatensatzes (SDS)

➢ Schrittweises und systematisches Erkennen von PIF in einem exemplarischen SDS aus strukturierten Routinedaten einer hausärztlichen Praxis

➢ Ergebnisbewertung im Sinne einer Datenschutz-Folgenabschätzung

Ziel: Erkennen von potentiell identifizierenden Feldinhalten (PIF)

- Untersuchung auf der Ebene

  (1) der Feldkennungen (Variablen, Attribute),

  (2) ihrer Kombinationen,

  (3) ihrer Feldinhalte (Ausprägungen, Werte), und

  (4) des gesamten Datensatzes

- Ergebnisbewertung im Sinne einer Datenschutz-Folgenabschätzung (DSFA)
  - Schwere eines möglichen Schadens
  - Eintrittswahrscheinlichkeit

Instrumente

für (1) und (2)

- Feldtyp
  Relative Häufigkeiten
  Semantische Gruppen
  Hausärztliche Expertise

für (3)

- TextCrawler ®
  Kategorien von Identifikatoren

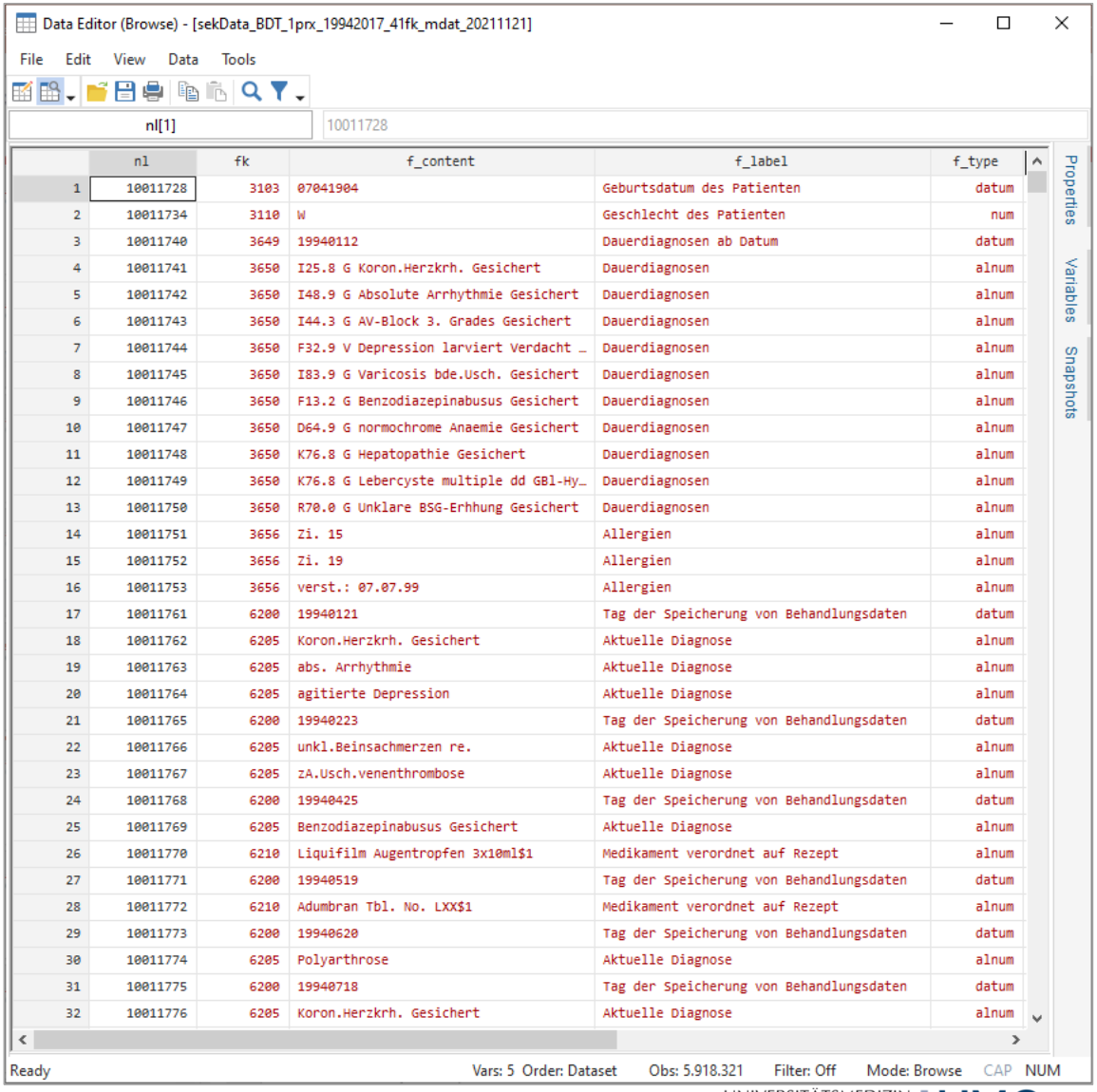
für (4)

- ARX ®

Sekundärdatensatz (SDS)

- 1 Hausarztpraxis
- N = 14.285 Patienten
- 01.01.1994 bis 31.12.2017
- Datenauswahl
  „RADAR“ AND „mdat“
- BDT-Variablen (fk): M_1 = 40
  40 aus dem RADAR Projekt
- 5.918.321 Datenzeilen

	f_type	_freq	ratio
1.	alnum	4831268	0.816
2.	datum	623393	0.105
3.	num	463660	0.078



	nl	fk	f_content	f_label	f_type
1	10011728	3103	07041904	Geburtsdatum des Patienten	datum
2	10011734	3110	W	Geschlecht des Patienten	num
3	10011740	3649	19940112	Dauerdiagnosen ab Datum	datum
4	10011741	3650	I25.8 G Koron.Herzkrh. Gesichert	Dauerdiagnosen	alnum
5	10011742	3650	I48.9 G Absolute Arrhythmie Gesichert	Dauerdiagnosen	alnum
6	10011743	3650	I44.3 G AV-Block 3. Grades Gesichert	Dauerdiagnosen	alnum
7	10011744	3650	F32.9 V Depression larviert Verdacht …	Dauerdiagnosen	alnum
8	10011745	3650	I83.9 G Varicosis bde.Usch. Gesichert	Dauerdiagnosen	alnum
9	10011746	3650	F13.2 G Benzodiazepinabusus Gesichert	Dauerdiagnosen	alnum
10	10011747	3650	D64.9 G normochrome Anaemie Gesichert	Dauerdiagnosen	alnum
11	10011748	3650	K76.8 G Hepatopathie Gesichert	Dauerdiagnosen	alnum
12	10011749	3650	K76.8 G Lebercyste multiple dd GBl-Hy…	Dauerdiagnosen	alnum
13	10011750	3650	R70.0 G Unklare BSG-Erhhung Gesichert	Dauerdiagnosen	alnum
14	10011751	3656	Zi. 15	Allergien	alnum
15	10011752	3656	Zi. 19	Allergien	alnum
16	10011753	3656	verst.: 07.07.99	Allergien	alnum
17	10011761	6200	19940121	Tag der Speicherung von Behandlungsdaten	datum
18	10011762	6205	Koron.Herzkrh. Gesichert	Aktuelle Diagnose	alnum
19	10011763	6205	abs. Arrhythmie	Aktuelle Diagnose	alnum
20	10011764	6205	agitierte Depression	Aktuelle Diagnose	alnum
21	10011765	6200	19940223	Tag der Speicherung von Behandlungsdaten	datum
22	10011766	6205	unkl.Beinsachmerzen re.	Aktuelle Diagnose	alnum
23	10011767	6205	zA.Usch.venenthrombose	Aktuelle Diagnose	alnum
24	10011768	6200	19940425	Tag der Speicherung von Behandlungsdaten	datum
25	10011769	6205	Benzodiazepinabusus Gesichert	Aktuelle Diagnose	alnum
26	10011770	6210	Liquifilm Augentropfen 3x10ml$1	Medikament verordnet auf Rezept	alnum
27	10011771	6200	19940519	Tag der Speicherung von Behandlungsdaten	datum
28	10011772	6210	Adumbran Tbl. No. LXX$1	Medikament verordnet auf Rezept	alnum
29	10011773	6200	19940620	Tag der Speicherung von Behandlungsdaten	datum
30	10011774	6205	Polyarthrose	Aktuelle Diagnose	alnum
31	10011775	6200	19940718	Tag der Speicherung von Behandlungsdaten	datum
32	10011776	6205	Koron.Herzkrh. Gesichert	Aktuelle Diagnose	alnum

(1) Feldkennungen des SDS

- 1 Hausarztpraxis
- N = 14.285 Patient*innen
- 01.01.1994 bis 31.12.2017
- Datenauswahl
  „RADAR“ AND „mdat“
- BDT-Variablen (fk): M_1 = 40
  40 aus dem RADAR Projekt

	f_type	_freq	ratio
1.	alnum	28	0.700
2.	datum	5	0.125
3.	num	7	0.175

	fk	f_label	f_type	_freq
1	0202	Praxistyp	num	1
2	0204	Arztgruppe verbal	alnum	1
3	0206	PLZ Ort	alnum	1
4	0225	Anzahl Ärzte	num	1
5	3103	Geburtsdatum des Patienten	datum	143622
6	3110	Geschlecht des Patienten	num	143589
7	3649	Dauerdiagnosen ab Datum	datum	3883
8	3650	Dauerdiagnosen	alnum	17539
9	3651	Dauermedikamente ab Datum	datum	267
10	3652	Dauermedikamente	alnum	481
11	3656	Allergien	alnum	5525
12	4101	Quartal der Abrechnung	num	115229
13	4105	Geschäftsstelle	alnum	11630
14	4107	Abrechnungsart (Schein)	num	94964
15	4121	Gebührenordnung	num	108978
16	5000	Leistungstag	datum	65780
17	5001	GNR/GNR-Ident	alnum	167077
18	6000	Abrechnungsdiagnose	alnum	168107
19	6001	ICD-Schlüssel	alnum	45033
20	6200	Tag der Speicherung von Behandlungsdaten	datum	409841
21	6205	Aktuelle Diagnose	alnum	778177
22	6210	Medikament verordnet auf Rezept	alnum	339751
23	6211	Außerhalb Rezept verordnetes Medikament	alnum	1
24	6215	Ärztemuster	alnum	1
25	6220	Befund	alnum	757482
26	6221	Fremdbefund	alnum	1
27	6222	Laborbefund	alnum	1
28	6225	Röntgenbefund	alnum	1
29	6260	Therapie	alnum	1
30	6265	Physikalische Therapie	alnum	1
31	6280	Überweisung Inhalt	alnum	3942
32	6285	AU Dauer	num	898
33	6286	AU wegen	alnum	2406
34	6290	Krankenhauseinweisung, Krankenhaus	alnum	1
35	6291	Krankenhauseinweisung wegen	alnum	1
36	8401	Befundart	alnum	509673
37	8410	Test-Ident	alnum	521252
38	8411	Testbezeichnung	alnum	521244
39	8420	Ergebnis-Wert	alnum	500569
40	8421	Einheit	alnum	481369

Properties Variables Snapshots

Ready Length: 10 Vars: 4 Order: Dataset Obs: 40 Filter: Off Mode: Browse CAP NUM

40 ausgewählte BDT-Felder, 11 semantische Gruppen

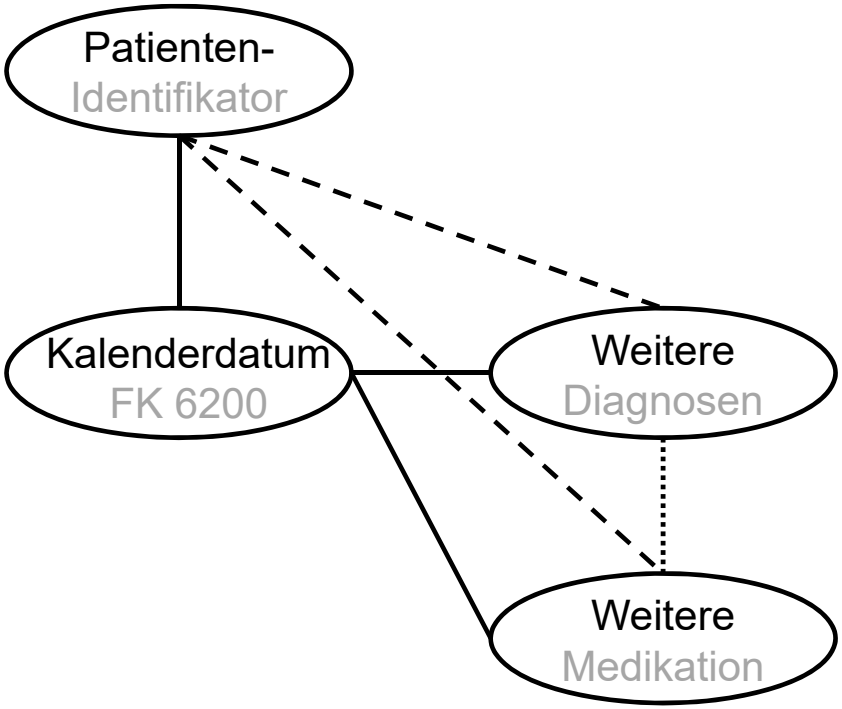
	Anzahl BDT-Felder	Semantische Gruppe	Beispiele	Risiko der Re-Identifizierung	Wichtigkeit für die Forschung
1	4	Diagnosen	Beratungsanlass; ICD-10	hoch	hoch
2	4	Medikation	Laufende Medikation	mittel	hoch
3	5	Laborergebnisse	Hb; BZ Kreat;	niedrig	mittel
4	4	Befunde	RR; BMI; Notizen	niedrig / hoch	mittel
5	2	Therapie	Physiotherapie	niedrig	mittel
6	5	Weitere Prozeduren	AU; Überweisung	niedrig	mittel
7	5	Zeit- und Datumsdaten	Einlesedatum	hoch	hoch
8	3	Stamm- und Dauerdaten des Patienten	Geschlecht, Geburtsjahr	hoch	hoch
9	4	Kenndaten der Praxis	Anzahl Ärzte	hoch (teilweise)	mittel
10	3	Kostenträger	GKV / Privat	niedrig	niedrig
11	1	Abrechnung	GOP	niedrig	mittel

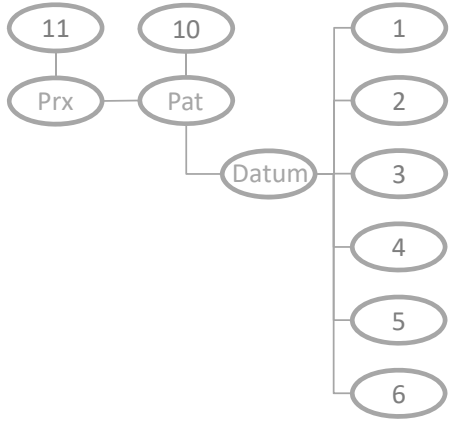
(1) Besonders kritische BDT-Feldkennungen

→	3656	Dauerbemerkungen	kritische Freitexteinträge
	6205	Aktuelle Diagnose	semantische Fehleinträge
→	6210	Medikament auf Rezept	semantische Fehleinträge
→	6220	Befund	kritische Freitexteinträge
	6225	Röntgenbefund	semantische Fehleinträge
	0204	PLZ Praxisort	→ Verzicht und Entfernung → IDAT
	3103	Geburtsdatum des Patienten	→ Geburtsjahr
	3110	Geschlecht des Patienten	unverändert

(2) Kombinationen von Feldkennungen

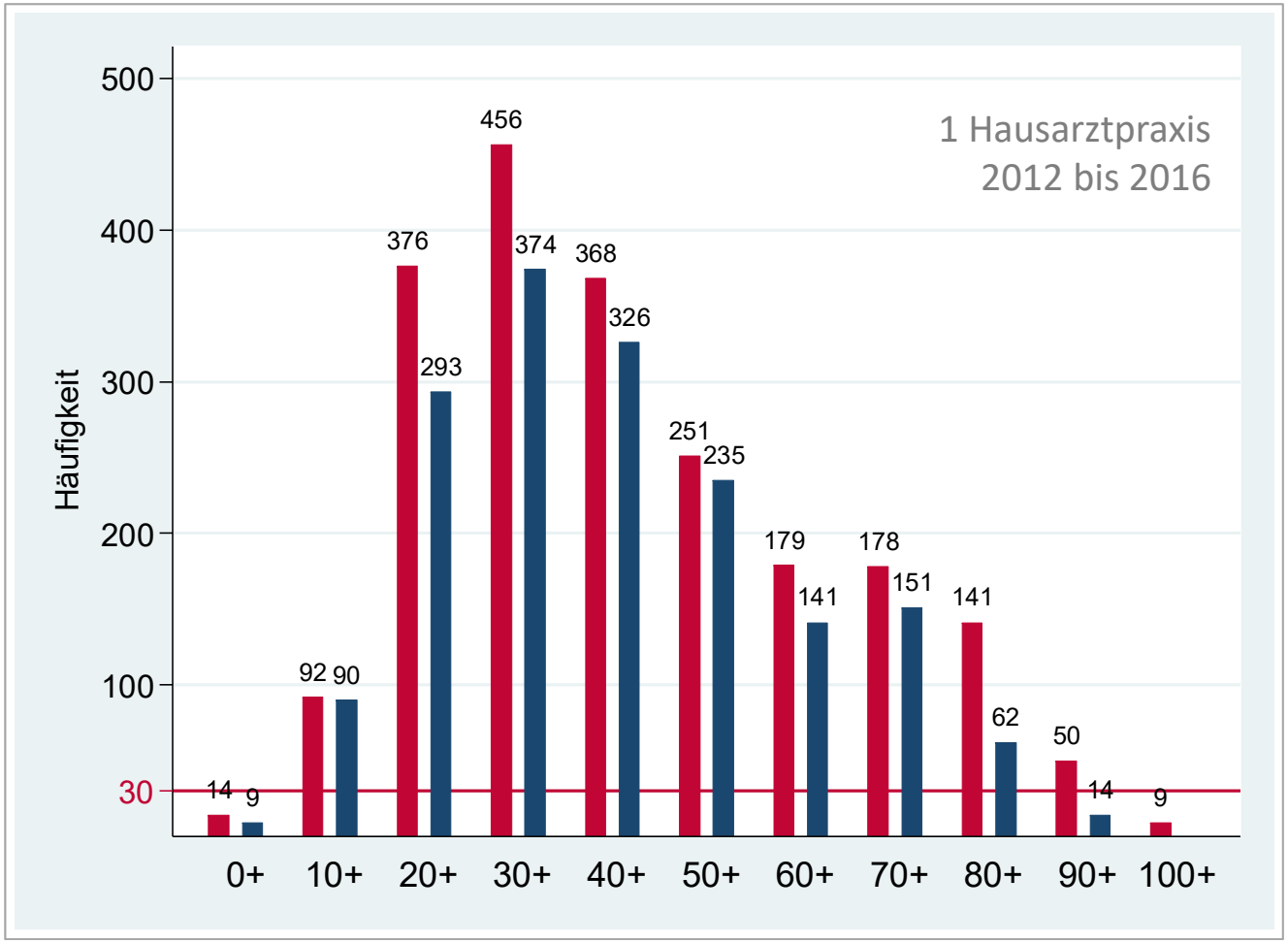
- Duplets
- Triplets
- Quadruplets
- Multiplets





(2) Kombinationen von Feldkennungen und Feldinhalten

- Altersdekade
  - Geschlecht



(3) Feldinhalte (Ausprägungen, Werte)

$$\hat{p} = \frac{n}{N}$$

Namen

- 10.451 Arztnamen
- 1.000 häufige Nachnamen
- 2.054 Städtenamen in Deutschland

	Quelldatensatz (SDS) 1		Quelldatensatz (SDS) 2	
	Anzahl Datenzeilen	*Wahrscheinlichkeit (Schätzer)*	Anzahl Datenzeilen	*Wahrscheinlichkeit (Schätzer)*
	n	***p̂***	**n**	***p̂***
Quelldatensatz (SDS)	5918321	*1,0000000*	340082	*1,0000000*
Schlüsseldaten "Ärzte"	10451		10451	
Treffer, gesamt	104025	*0,0175768*	5499	*0,0161696*
Namen				
Treffer, richtig-positiv	359	*0,0000607*	191	*0,0005616*
Namen, falsch-negativ	1	*0,0000002*	1	*0,0000029*
Arztnamen				
Treffer, richtig-positiv	68	*0,0000115*	176	*0,0005175*
Namen, richtig-positiv	29	*0,0000049*	36	*0,0001059*
Schlüsseldaten "Namen"	1000		1000	
Treffer, gesamt	31140	*0,0052616*	2625	*0,0077187*
Namen				
Treffer, richtig-positiv	254	*0,0000429*	76	*0,0002235*
Namen, richtig-positiv	75	*0,0000127*	25	*0,0000735*
Schlüsseldaten "Städte"	2054		2054	
Treffer, gesamt	15227	*0,0025729*	234	*0,0006881*
Städtenamen				
Treffer, richtig-positiv	(many)		(many)	
Namen, richtig-positiv	13	*0,0000022*	19	*0,0000559*
Namen				
Treffer, richtig-positiv	34	*0,0000057*	2	*0,0000059*
Namen, richtig-positiv	4	*0,0000007*	2	*0,0000059*

(3) Feldinhalte (Ausprägungen, Werte)

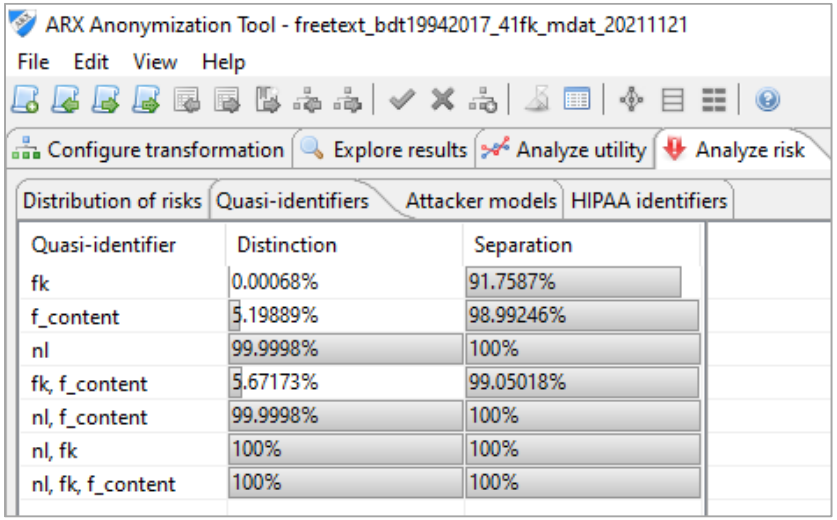
$$\hat{p} = \frac{n}{N}$$

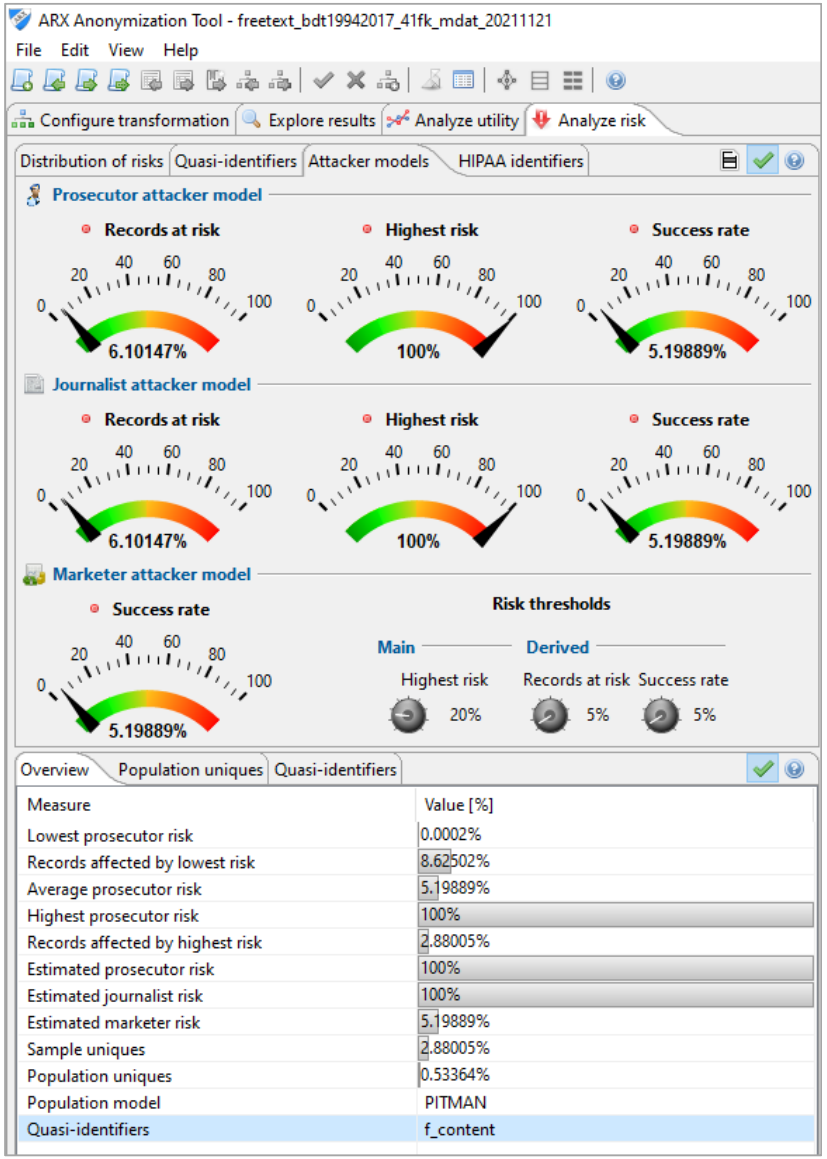
- PLZ
- Telefonnummern

	Quelldatensatz (SDS) 1		Quelldatensatz (SDS) 2	
	Anzahl Datenzeilen	*Wahrscheinlichkeit (Schätzer)*	Anzahl Datenzeilen	*Wahrscheinlichkeit (Schätzer)*
	n	***p̂***	**n**	***p̂***
Quelldatensatz (SDS) Schlüsseldatei "PLZ TelNr"	5918321 (Befehle)	*1,0000000*	340082 (Befehle)	*1,0000000*
Treffer, gesamt	2553	*0,0004314*	9	*0,0000265*
PLZ TelNr				
Treffer, richtig-positiv	0	*0,0000000*	8	*0,0000235*
Entität, richtig-positiv	0	*0,0000000*	1	*0,0000029*
Entität, falsch negativ	1	*0,0000002*	0	*0,0000000*
TelNr				
Treffer, richtig-positiv	0	*0,000000*	36	*0,000106*
Entität, richtig-positiv	0	*0,000000*	31	*0,000091*
Entität, falsch negativ	0	*0,000000*	0	*0,000000*

ARX Risikoanalyse: Quasi-Identifikatoren &

- Sekundärdatensatz (SDS)
- 5.918.321 Datenzeilen



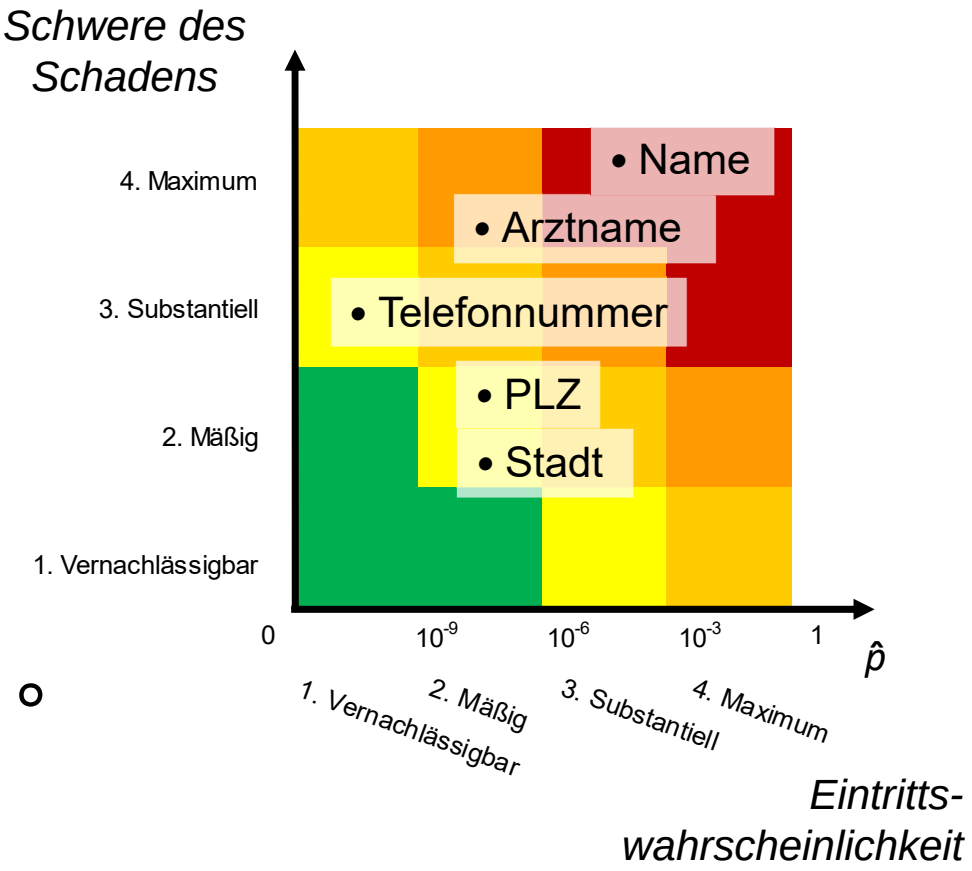


3 Angreifer-modelle

Ergebnisbewertung

- *Privacy by design*
  Datensparsamkeit / -minimierung
- Anpassungen

- Potentiell identifizierende Feldinhalte
- Datenschutzfolgenabschätzung
  DSFA nach Art. 35 DSGVO
  - Schwere des möglichen Schadens
  - Eintrittswahrscheinlichkeit



Diskussion und Schlussfolgerungen

- Untersuchungen zu PIF müssen immer 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.
- 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
- Eine semantische Strukturierung der Daten, etwa unter SNOMED CT, ist erstrebenswert, hilft jedoch PIF durch Fehleingaben nicht ab

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. Secondary Use of Electronic Medical Record Data from Primary Health Care is Feasible: Report from RADAR Project. Gesundheitswesen 83, 1–9 (im Druck)

[2] TextCrawler Free 3.0, 2014, DigitalVolcano Software

[3] ARX Data Anonymization Tool; https://arx.deidentifier.org/ Zugriff 23.11.2021

Abstrakt für AGENS Methodenworkshop, Mai 2022

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 (1) der Feld-kennungen (Variablen, Attribute), (2) ihrer Kombinationen, (3) ihrer Feldinhalte (Ausprägungen, Werte) und (4) des gesamten Datensatzes. Instrumente sind für (1) und (2) Feldtyp, relative Häufigkeiten, Kategorien, und hausärztliche Expertise, (3) TextCrawler [2], (4) 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“, „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, pseudo-nymisierter 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.

Routine Anonymized Data for
Advanced Health Services Research