**Appendix**

**Re-Identifizierungspotential in 40 ausgewählten Variablen (Feldern) des BDT-Routinedatensatzes einer Arztpraxis**

Im Folgenden werden die zu 11 semantischen Gruppen zusammengefassten 40 BDT-Variablen anhand der drei Kriterien

**1** Expertenwissen

**2** Recherchierbares Zusatzwissen

**3** Statistische Häufigkeit

unter dem Aspekt der Re-Identifizierung ihrer Träger (betroffene Personen) detailliert untersucht und bewertet.

*Gruppe 1, Diagnosen*

entspricht Basismodul 4, „Diagnosen“, Kerndatensatz der M-Initiative [24]

Hier interessierten vor allem die Variablen (BDT-Felder) 3650 (Dauerdiagnose), 6000 (Abrechnungsdiagnose) und 6205 (Aktuelle Diagnose; Informationen zu Krankheitsbezeichnungen oder Beratungsanlässen), alle drei in Form von Langtextfelder mit bis zu 70 Zeichen, bzw. 6001 (Diagnosen als endständige ICD-Kodes).

**1** Medizinprofessionelle erkennen unter den 30 häufigsten Diagnosen [21] leicht „Adipositas“ (ICD-10 E66) sowie ggf. „Schädlicher Tabakkonsum“ (F17) am Erscheinungsbild. Sie erkennen auch: Akute Infektionen ... der oberen Atemwege (J06), Sonstige chronische obstruktive Lungenkrankheit (J66), Asthma bronchiale (J45), Gonarthrose [Arthrose des Kniegelenkes; (M17)], Senilität (R54), aber auch weitere sieben Diagnosen (M54: Rückenschmerzen, I25: Chronische ischämische Herzkrankheit, I50: Herzinsuffizienz). Medizinprofessionelle erkennen unter besonderen Umständen auch weitere Diagnosen (F32: Depressive Episode, I83: Varizen der unteren Extremitäten, M47: Spondylose, J30 Vasomotorische und allergische Rhinopathie).

Phänotypisch leicht erkennbar sind beispielsweise bestimmte genetische Ausprägungen (z.B. Trisomie 21, „Down-Syndrom“, oder Kleinwuchs), Fehlen von Extremitäten (Amputationen, kongenitale Anlagestörungen), Versorgung mit Körperersatzteilen oder Prothesen (Brille; „Glasauge“).

**2** Recherchierbares Zusatzwissen über „Diagnosen“ ist mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung leicht und in großer Detailtiefe zu erlangen und gut verknüpfbar.

**3** Die Streuung aller berichteten ICD-Codes in Praxen ist stark [32]: 18 bis 91% aller Diagnosen einer Praxis wurden weniger als fünfmal beobachtet; Median: 46%; MW: 47% (SD 15%); IQR 34% - 57%.

*Gruppe 2, Medikation*

(entspricht Basismodul 7, „Medikation“, z. T. Basismodul 5, „Prozeduren“ im Kerndatensatz der M-I)

Variablen (Feldkennungen): 3652 (Dauermedikamente), 6210 (Medikament verordnet auf Rezept), 6211 (Auß­erhalb Rezept verordnetes Medikament) und 6215 (Ärztemuster).

**1** Für Professionelle sind vermutlich 11 unter den 76 häufigsten ATC-Kodes [23] leicht erkennbar und einem Krankheitsbild/Diagnose zuordenbar, bei den allermeisten der ATC-Kodes ist ihnen ein Rückschluss auf die behandelte Krankheit oder zumindest eine kleine Zahl von Krankheiten möglich. Aus zusätzlicher zufälliger Beobachtung einer Index-Person wird der Professionelle jedoch nur selten oder gar nicht eine Medikamentenzuordnung treffen können. Gleichzeitiges Auftreten mehrerer Ausprägungen (Kombination; Multimedikation) erhöht das Re-Identifizierungsrisiko geringfügig.

Spezifische Ausprägungen, die eine Medikation sicher erkennen und so einer Person zuordnen lassen, sind beispielsweise parenteral zu verabreichende Medikamente wie Insulininjektionen, Atemwegstherapeutika wie Inhalatoren, oder bestimmte Medikamentenzubereitungen oder -kombinationen.

**2** Recherchierbares Zusatzwissen über „Medikation“ ist mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung leicht zu erlangen und verknüpfbar.

**3** Unter allen ATC-Dreistellern in 160 Praxen [32] liegt der prozentualer Anteil diejenigen, die je Praxis weniger als fünfmal beobachtet werden, zwischen 0% und 100%. Median: 12%; MW: 28% (SD 32%; IQR 7% - 50%)

*Gruppe 3, Laborergebnisse*

(entspricht Basismodul 6, „Laborbefunde“ im Kerndatensatz der M-I)

Laborergebnisse eines Patienten werden in BDT unter den Variablen (Feldkennungen) 8401 (Befundart), 8410 (Test-Ident), 8411 (Testbezeichnung), 8420 (Ergebnis), 8421 (Einheit) abgelegt.

**1** Der Professionelle kann zwar bei vielen Laborwerten und -ergebnissen auf die Krankheit oder zumindest eine kleine Zahl oder Untergruppe von Krankheiten schließen. Aus zusätzlicher zufälliger Beobachtung einer Index-Person wird er jedoch nur selten oder gar nicht eine Zuordnung treffen können. Das gleichzeitige Auftreten mehrerer Laborwerte (Kombination), insbesondere mit einem Abnahmedatum oder dem Tag der Speicherung von Ergebnissen, erhöht das Re-Identifizierungsrisiko.

Einzelne Labordatenergebnisse, die phänotypisch erkennbar sein können, sind beispielsweise deutlich erhöhte Bilirubinwerte („Gelbsucht“, Ikterus) oder etwa erhöhte Serum-Kupferwerte, die in Verbindung mit einem „Kayser-Fleischer-Kornealring“, einem braunen Ring durch Kupfereinlagerungen am Rand der Iris, pathognomonisch für eine seltene Kupferspeicherkrankheit (Morbus Wilson) sind.

**2** Recherchierbares Zusatzwissen über „Medikation“ ist mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung leicht zu erlangen und verknüpfbar.

**3** Nur wenige Laborergebnisse, nämlich nur bei hausärztlich selten untersuchten Parametern, werden weniger als fünfmal je Praxis angeführt [32].

*Gruppe 4, Befunde*

(entspricht dem Basismodul 5, „Prozeduren“ im Kerndatensatz der M-I)

Befunde werden unter den Variablen (Feldkennungen) 6220 (Befund), 6221 (Fremdbefund), 6222 (Laborbefund) und 6225 (Röntgenbefund) abgelegt.

**1** Dem medizinprofessionellen Beobachter werden Befunde insbesondere in Kombination mit weiteren Informationen (Diagnosen, Medikamente) eine Re-Identifizierung in einigen Fällen erlauben (beispielsweise Wernicke-Mann-Gang nach Schlaganfall, Parkinsongang, „Steppergang“ durch Lähmung eines Fußhebermuskels).

Zusätzlich gilt das allgemeine erhebliche Risikopotential von Freitexteinträgen. Hier dokumentieren Ärzte und Medizinische Fachangestellte Patientenbefunde der Untersuchung, aber noch viel mehr: so auch Behandlungsanlässe, die oftmals in einem spezifischen psychosozialen Kontext stehen, auch Termine bei Mitbehandlern, Arztnamen der aktuell behandelnden Ärzte in Praxen mit mehreren Ärzten; Namen von Verwandten, Bezugspersonen oder Dritten. Besonders schutzwürdig sind Angaben zu beruflichen und privaten Konfliktthemen, Sexualanamnese, Strafsachverhalten u.v.m.

**2** Recherchierbares Zusatzwissen über „Befunde“ kann mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung erlangt werden.

**3** Da die Informationen zumeist als strukturierter Freitext vorliegen, erlauben Häufigkeitsanalysen zumeist keine Re-Identifizierung.

*Gruppe 5, Therapien*

(entspricht dem Basismodul 5, „Prozeduren“ im Kerndatensatz der M-I)

Daten finden sich in den Variablen (Feldern) 6260 (Therapie) und 6265 (Physikalische Therapie).

**1** Der Professionelle wird angegebene Therapien, insbesondere kombiniert mit weiteren Informationen (Diagnosen, Medikamente), zur Re-Identifizierung in einigen Fällen nutzen können.

Es gilt das allgemeine Risikopotential von Freitexteinträgen, analog zur entsprechenden Anmerkung in Gruppe 4.

**2** Recherchierbares Zusatzwissen über „Therapien“ kann teilweise mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung erlangt werden.

**3** Da die Informationen zumeist als strukturierter Freitext vorliegen, erlauben Häufigkeitsanalysen zumeist keine Re-Identifizierung.

*Gruppe 6, Weitere Prozeduren*

(entspricht dem Basismodul 5, „Prozeduren“ im Kerndatensatz der M-I)

Hier finden sich Prozeduren im engeren Sinne in den Variablen (Feldern) 6280 (Überweisung), 6285 (Arbeitsunfähigkeitsbescheinigung Dauer) und 6286 (Arbeitsunfähigkeitsbescheinigung wegen …), 6290 (Krankenhauseinweisung) und 6291 (Krankenhauseinweisung wegen …).

**1** Der medizinisch Professionelle wird eine Kenntnis derartiger Prozeduren nur zufällig mit einem beobachteten Fernbleiben vom Arbeitsplatz oder Aufenthalt im Krankenhaus der Index-Person verknüpfen können und damit kaum zusätzliche Re-Identifizierungsmöglichkeit haben.
Eine Überweisung beispielsweise zum Hämatologen/Onkologen oder zur Genetischen Beratung kann in Kombination mit anderen Variablen oder unter besonderen Umständen auch phänotypisch re-identifizierend sein.

**2** Recherchierbares Zusatzwissen über „Prozeduren“ bei Index-Personen kann vermutlich nur selten mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung erlangt werden.

**3** Die Prozeduren „Überweisung“ und „Arbeitsunfähigkeit“ kommen in der Hausarztpraxis häufig vor (Re-Identifizierungspotential gering); „Krankenhauseinweisungen“ hingegen sind selten und können deshalb, insbesondere in Kombination mit anderen Informationen (z. B. Datum) identifizierend sein, insbesondere wenn die absolute Häufigkeit klein ist (beispielsweise *k* < 5).

*Gruppe 7, Zeit- und Datumsdaten*

(entspricht Basismodul 3, „Falldaten“ im Kerndatensatz der M-I)

Ein Kalenderdatum erscheint als Variableninhalt unter der Feldkennung 3649 (Dauerdiagnosen ab Datum), 3651 (Dauermedikamente ab Datum), 5000 (Leistungstag) und 6200 (Tag der Speicherung von Behandlungsdaten), eine Angabe zum laufenden Quartal unter 4101 (Quartal der Abrechnung). Die Besonderheit von Zeit- und Datumsdaten liegt in ihrer einzigartigen Auswirkung bei regelhafter Kombination mit anderen Kalender- oder anderen Informationen, woraus dann auf Reihenfolgen, gemeinsames Auftreten, Verknüpfungen sowie Auslöser/Folgen, gar Ursache/Wirkungen geschlossen werden kann.

**1** Einzelne Datumsangaben für sich allein haben auch für den medizinisch geschulten Beobachter kein wesentliches Re-Identifizierungsrisiko, in ihrer Reihung jedoch sehr wohl.
Nur selten werden einzelne Datumsangaben in Verbindung mit anderen allgemein verfügbaren Informationen (Großereignisse, Massenphänomene) Bedeutung erlangen.

**2** Zeitangaben führen bei entsprechend hoher zeitlicher Auflösung schnell zu hoch individuellen Wertekombinationen (vergl. auch Kriterium 3), die eine externe Zuordnung von Datensätzen stark vereinfacht, zumal in Bezug auf die zeitlichen Angaben zumeist mit externem Zusatzwissen gerechnet werden muss.

**3** Eine einzelne Datumsangabe für sich lässt selten eine Person erkennen, in ihrer Häufung jedoch bergen Kalenderdaten sehr wohl ein Re-Identifizierungspotential. Bei hoher absoluter Häufigkeit von Datumsangaben kann vermutet werden, dass das Individuum einen großen hausärztlichen Versorgungsbedarf hat (siehe auch [22]).

*Gruppe 8, Stamm- und Dauerdaten des Patienten*

(entspricht Basismodul 2, „Demographie“ im Kerndatensatz der Medizininformatik-Initiative)

Variablen (Felder): 3110 (Geschlecht), 3103 (Geburtstag / Geburtsjahr / Alter in Jahren), 3656 (Dauerbemerkung / „Allergie“).

**1** Bei hausärztlichen Patienten taugen Angaben zum Geschlecht kaum zur Re-Identifizierung. Die Kenntnis des Geburtstages einer Person hingegen ist generell als quasi-identifizierend anzusehen. Deshalb sollte in Routinedaten-Projekten – ob mit und ohne Einwilligung der Betroffenen – von vornherein allein das Alter in Jahren verwendet werden. Medizinprofessionelle ziehen vermutlich aus Alters- und Geschlechtsangaben eines Individuums insgesamt mehr rahmende und einordnende Informationen als Laien.
Für „Dauerbemerkungen“ gilt das allgemeine Risiko von Freitexteinträgen. Unter Dauerbemerkungen können auch phänotypisch wahrnehmbare Patienteneigenschaften („schwerhörig“), Berufsbezeichnungen oder die Teilnahme an der Hausarztzentrierten Versorgung oder an Strukturierten Behandlungsprogrammen (DMP) ausgeprägt bzw. vermerkt sein.

**2** Recherchierbares Zusatzwissen über „Geschlecht“ und auch über das Alter einer Index-Person ist mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung sehr leicht zu erlangen und verknüpfbar.

**3** In der Hausarztpraxis ist die Geschlechtsverteilung annähernd gleich der generellen Bevölkerung. Sehr junge und sehr alte Patienten kommen praxisspezifisch möglicherweise in Häufigkeiten vor, die eine zuvor gesetzte Grenze (beispielsweise *k* = 5) unterschreiten.

*Gruppe 9, Kenndaten der Praxis*

(entspricht z. T. dem Basismodul 1, „Person“ im Kerndatensatz der M-I)

Variablen (Felder): 0202 (Praxistyp), 0204 (Arztgruppe verbal), 0206 (Postleitzahl und Ort der Praxis), 0225 (Anzahl der Ärzte)

**1** Der Professionelle erhält aus dem Inhalt der Variablen (Feldkennung) 0206 eine Information, die annähernd wie ein direkter Identifikator, also als Quasi-Identifikator einer Index-Person anzusehen ist. In Kombination damit machen auch die übrigen Felder dieser Gruppe die Re-Identifizierung der Hausarztpraxis und damit der Praxismitarbeiter möglich. Unverarbeitete PLZ und Ortsangaben auch der betreuenden oder arbeitgebenden Hausarztpraxis erhöhen das Re-Identifizierungsrisiko für einen Betroffenen erheblich. Standortdaten werden in Artikel 4 (1) DSGVO explizit zu den personenbezogenen Daten gezählt.
In Regionen mit geringer Hausarztdichte erhöht die Kenntnis der PLZ das Re-Identifizierungspotential einer Praxis und der mit ihr verbundenen Personen. Orts- und Geodaten werden an zweiter Stelle der 18 HIPAA-Identifikatoren genannt. „*All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code, if according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000*“ [19].

**2** Recherchierbares Zusatzwissen aus der PLZ einer Praxis ist mit den bestehenden Möglichkeiten elektronischer Datenverarbeitung leicht zu erlangen und verknüpfbar.

**3** Häufigkeitsanalysen des Praxistyps sind unproblematisch, weil annähernd gleich verteilt [33]. Die Arztgruppen „Allgemeinarzt und Praktischer Arzt“, „Internist“ und „Kinderarzt“ sind häufig und Angaben insofern unproblematisch bezüglich ihres Re-Identifizierungspotentials. Angaben über Anzahl der mitarbeitenden Ärzte und Ärztinnen lassen nur bei hohen Ausprägungen eine schwere bis mittelschwere Identifizierung der Praxis zu.

*Gruppe 10, Kostenträger*

(entspricht teilweise dem Basismodul 1, „Person“ im Kerndatensatz der M-I)

Variablen (Feldkennungen) 4105 (Geschäftsstelle), 4107 (Abrechnungsart (Schein)), 4121 (Gebührenordnung)

**1** Für den Beobachter mit medizinischem Erfahrungshintergrund haben Informationen über den Kostenträger kein wesentliches Potential, eine beobachtete Index-Person zu identifizieren.

**2** Zusatzwissen zu Kostenträgern ist leicht recherchierbar.

**3** Kenntnis über die genutzte „Gebührenordnung für Ärzte (GOÄ)“ erlaubt Identifizierung des Patienten als Selbstzahler oder Privatpatient und damit Zuordnung zu 10% der Bevölkerung der Bundesrepublik.

*Gruppe 11, Abrechnung*

(entspricht dem Erweiterungsmodul 16, „Entgelte“ im Kerndatensatz der M-I)

Variable (Feldkennung) 5001 (GNR / GNR-Ident)

**1** Der Medizinprofessionelle kann aus einer kleinen Zahl von bestimmten Variableninhalten mit der abgerechneten Leistung auf die behandelte oder auszuschließende Krankheit rückschließen und dies möglicherweise mit Informationen aus anderen Variablen einer beobachteten Person in Verbindung bringen.
Vereinzelt sind erbrachte hausärztliche Leistungen ohne weiteres zu erkennen (Beispiel: Wundverband).

**2** Zusatzwissen über die Gebührenordnungsnummern (GNR) in der Abrechnung ambulanter medizinischer Leistungen kann leicht erlangt und verknüpft werden.

**3** Eine große Anzahl abgerechneter Leistungen bei einem Patienten lässt vermuten, dass das Individuum einen großen hausärztlichen Versorgungsbedarf hat. Ziffern ausschließlich oder überwiegend aus der GOÄ identifizieren den Patienten als Privatpatienten oder Selbstzahler (s. **3** in der Gruppe „Kostenträger“)**.** Bereits durch den Ansatz einer einzigen GNR wird ein Individuum als Nutzer ambulanter Versorgung und des Gesundheitssystems erkennbar und damit zum Patienten.