Datenschutz in der medizinischen Forschung

GMDS-Jahrestagung Mainz, 28. September 2011

Univ.-Prof. Dr. Klaus Pommerening
Universitätsmedizin Mainz, IMBEI

Dr. Johannes Drepper
TMF e. V. Berlin

Gefördert vom
Bundesministerium für Bildung und Forschung

Inhalt

1. Szenario und Anwendungsfälle
2. Rechtliche Grundlagen der medizinischen Forschung
3. **Anonymisierung und Pseudonymisierung**
4. Patienteninformation und Einwilligungserklärung
5. Die Datenschutzkonzepte der TMF
6. Praktisches Vorgehen

Anonymisierung (Definition)

BDSG §3 (6): ***Anonymisieren*** *ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem* unverhältnismäßig großen Aufwand *an Zeit, Kosten und Arbeitskraft* einer bestimmten oder bestimmbaren natürlichen Person zugeordnet *werden können.*

[„Faktische Anonymisierung“:
d. h., das Prinzip der Verhältnismäßigkeit wird berücksichtigt.]

Ziel der Anonymisierung: Daten nicht mehr personenbeziehbar.

- (Wirksam) *Anonymisierte Daten sind nicht den Datenschutzregeln unterworfen*.

Aspekte der Anonymisierung

Personenbeziehbarkeit hängt ab vom möglichen Zusatzwissen und sonstigen Möglichkeiten des Empfängers,

- und ist daher im Einzelfall zu beurteilen.

Wirksamkeit der Anonymisierung muss **bewertet** werden (z. B. bei Weitergabe von Daten oder Proben)

- im Einzelfall
- und **immer wieder neu**
- nach dem aktuellen Stand der Technik
- und im Hinblick auf die Möglichkeiten der Empfänger.

Anonymisierung betrifft Datenempfänger als natürliche Person,

- bei multizentrischen Studien mit Konsiliartätigkeit unnötig gegenüber Studienleiter, der den Patienten sowieso kennt,
- aber nicht gegenüber dessen Forschungspersonal.

Reidentifizierungsrisiko

Das Reidentifizierungsrisiko (RI-Risiko) beschreibt die Wirksamkeit der Anonymisierung:

Wie hoch ist das Risiko, dass ein (weitergegebener) Datensatz einem Individuum zugeordnet werden kann?

Medizinische Forschung erzeugt hochdimensionale individuelle Datensätze RI-Risiko potenziell hoch.

„Externes“ Wissen nimmt zu und ist immer leichter beschaffbar:

- genetische Fingerabdrücke (deCODEme $500),
- assistierende Technik,
- soziale Netze und andere Internet-Aktivitäten (Bewegungsprofile, freiwillige Angaben über Tagesablauf, Medikamente, Suchanfragen, „intelligente“ Stromzähler, …).
- Schon die Uhrzeit einer einzelnen Arztvisite kann zur Identifizierung ausreichen.

Ausblick: „The Race for the 1000 $ Genome“:

- vollständige Sequenzierung

Beurteilung des Reidentifizierungsrisikos

Beurteilung oft nur im Einzelfall möglich.

Beispiel anonyme Übermittlung an Bundesstatistik (SchwbG* § 53):

- Alter, Geschlecht, Staatsangehörigkeit, Wohnort
- Art, Ursache und Grad der Behinderung

RI-Risiko bei Behinderten oft besonders hoch, da Behinderung „sichtbar“:

- physisch, Rollstuhl, Armbinde
- Schwerbehindertenausweis (SchwbG §4)
- Lohnsteuerkarte
- Parkausweis, Behindertenparkplatz
- Wohnung im Heim

[* Schwerbehinderten-Gesetz = SGB IX]

Wie anonymisiert man?

1. **Schritt** (nicht immer ausreichend): Entfernen von Identitätsdaten.
2. **Schritt** (oft notwendig): Vergröberung der „Nutzdaten“.

Entfernen von Identitätsdaten

- Namen, Adresse, Geburtsdatum

Entfernen nicht geheimer Identifikationsnummern

- Kontonummer, Versicherungsnummer, Fallnummer, ...

Entfernung oder Vergröberung typischer Einzelangaben

- Z. B. „Größe 1.42 m“ → „≤ 1.50 m“
- Art und Prozentsatz der Behinderung → „SB/GL*“
- Wohnort „87654 Einöd“ → „ländlich in 87???“

[* schwerbehindert/ gehörlos

Besonderheiten bei genetischen Daten

- Verfügbarkeit
- Unteilbarkeit (in Proben)
- Identifizierbarkeit

Genetische Daten sind nicht geheim: Referenzproben leicht zu gewinnen.

Genetische Daten sind Identifikatoren: 70 SNPs* genügen.

Bei seltenen Krankheiten reicht evtl. 1 SNP.

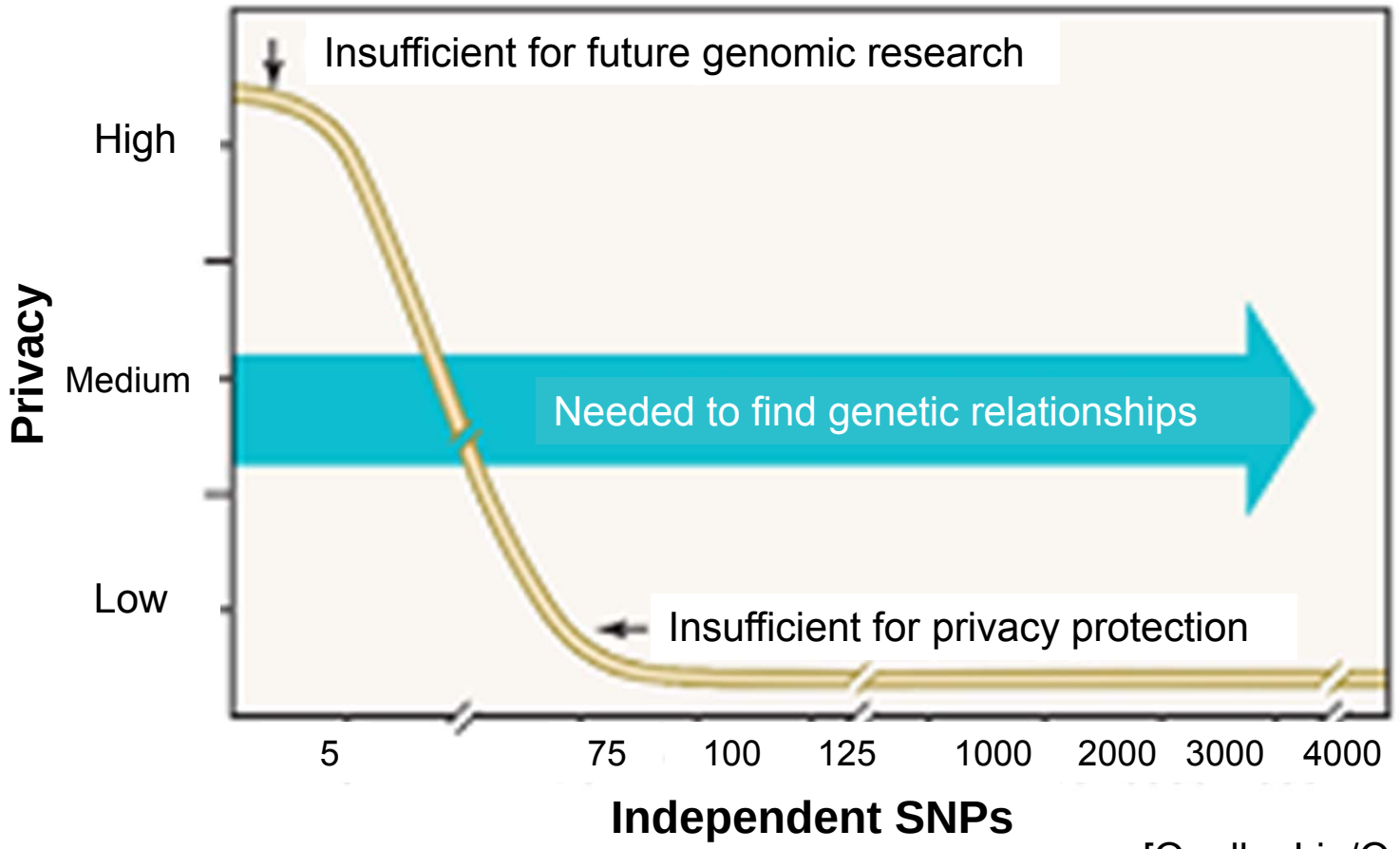
- Externes Wissen: Krankheit bekannt oder offensichtlich.

Für diagnostische und therapeutische Anwendungen typischerweise 1 bis 20 Gene mit ihren Varianten benötigt.

[* Single Nucleotide Polymorphism, Punktmutation]

Genetische Daten und Identifizierbarkeit

70 SNPs genügen zur Identifizierung.
Sinnvolle medizinische Anwendungen bewegen sich bereits in einem Datenschutz-kritischen Bereich.



[Quelle: Lin/Owen/Altman Science 305 (2004)]

k-Anonymität

Eine Datensammlung ist ***k*-anonym**,
wenn die Kombination der auch in anderen Datensammlungen vorhandenen Attribute in *k* verschiedenen Datensätzen innerhalb der Datensammlung vorkommt.

Als andere Datensammlungen sind die zu berücksichtigen,
die einem potenziellen Angreifer zum Abgleich zur Verfügung stehen, insbesondere öffentlich verfügbare oder soziodemographische Datenzusammenstellungen.

k-Anonymität bezieht sich immer auf einen festen Satz von Attributen.
Jede Ausprägung davon kommt $\geq k$-mal vor.

Beispiel: fiktiver Datensatz

Vorname	Geburtstag	Geschlecht	PLZ	Diagnose
Hans	17.04.75	M	76227	Impotenz
Peter	31.07.75	M	76228	Hodenkrebs
Karl	17.01.75	M	76227	Sterilität
Till	05.07.81	M	76133	Schizophrenie
Kai	31.12.81	M	76139	Diabetes
Lisa	05.07.83	W	76133	Magersucht
Nora	31.10.83	W	76131	Magersucht
Hans	17.04.75	M	76227	Blinddarmentzündung

Personen i. w. eindeutig, z. B. durch Vorname oder Geburtstag.
(Zu Hans gibt es zwei Datensätze.)
Angreifer mit Zugriff auf DB möchte Diagnose von Kai herausbekommen.

Beispiel: Formale Anonymisierung

Kennzahl	Geburtstag	Geschlecht	PLZ	Diagnose
123	**.**.75	M	76227	Impotenz
254	**.**.75	M	76228	Hodenkrebs
645	**.**.75	M	76227	Sterilität
459	**.**.81	M	76133	Schizophrenie
387	**.**.81	M	76139	Diabetes
240	**.**.83	W	76133	Magersucht
327	**.**.83	W	76131	Magersucht
123	**.**.75	M	76227	Blinddarmentzündung

Manche Personen immer noch eindeutig,
insb., wenn bekannt ist, dass sie in der Datensammlung sind.
- Kai (Adresse und Jahrgang bekannt) hat Diabetes.

Beispiel: 2-Anonymität

Kennzahl	Geburtstag	Geschlecht	PLZ	Diagnose
123	**.**.75	M	7622*	Impotenz
254	**.**.75	M	7622*	Hodenkrebs
645	**.**.75	M	7622*	Sterilität
459	**.**.81	M	7613*	Schizophrenie
387	**.**.81	M	7613*	Diabetes
240	**.**.83	W	7613*	Magersucht
327	**.**.83	W	7613*	Magersucht
123	**.**.75	M	7622*	Blinddarmentzündung

Jede Merkmalskombination aus
(Geburtstag, Geschlecht, PLZ) kommt mindestens 2x vor.
 Kai (Adresse und Jahrgang bekannt) könnte Schizophrenie oder Diabetes haben.

Szenarien

- Langfristspeicherung mit gelegentlichen (internen) Auswertungen
- Herausgabe von Daten zu einmaliger externer Auswertung (Scientific Use)
- Herausgabe von Daten zum freien Gebrauch (Public Use)

Musterhaft: Regelungen der Forschungsdatenzentren der statistischen Ämter des Bundes und der Länder. Siehe

http://www.forschungsdatenzentrum.de/datenzugang.asp

Nutzung von Forschungsdaten

„**Scientific Use**“: Export zur Auswertung/ Forschung nur mit zusätzlichem Anonymisierungs- oder Pseudonymisierungsschritt

- nach Abschätzung des RI-Risikos,
- *k*-Anonymisierung (z. B. *k* = 5)
- mit vertraglicher Vereinbarung, die Aufbewahrung oder Weiterverwendung ausschließt.
- Sonst Recherche-Aufträge durch interne Mitarbeiter ausführen.

„**Public-Use**“-Dateien mit medizinischen Daten kaum möglich, da Datensätze hochdimensional und Zusatzwissen beliebiger Nutzer nicht bekannt.

Also mit starker Vergröberung, wenn überhaupt!
RI-Risiko muss zuverlässig ausgeschlossen werden –
auch für die Zukunft und beliebige Weiterverbreitung.

Anonymisierung von Proben

Absolute Anonymisierung von Proben ist nicht möglich.

- Die Rückidentifizierung (über Referenzproben) ist noch aufwendig.
- Anonymisierbarkeit (de facto) z. Z. oft noch angenommen.

Anonymisierung ist keine ausreichende Grundlage für langfristige Aufbewahrung und Nutzung von Proben.

Eine geplanten Anonymisierung von Probenmaterialien ist dem Spender mitzuteilen.

- Der Patient ist darauf hinzuweisen, dass er bestimmte Rechte nach Anonymisierung nicht mehr wahrnehmen kann: Auskunft, Rückruf.

Anonymisierung von Proben verletzt das Persönlichkeitsrecht!

Für eine langfristige Nutzung von Biomaterialien sollte die Einwilligung in *pseudonymisierte* Lagerung und Verarbeitung vereinbart werden.

Das Persönlichkeitsrecht bleibt gewahrt, da die Proben wieder gefunden werden.

Pseudonymisierung (Definition)

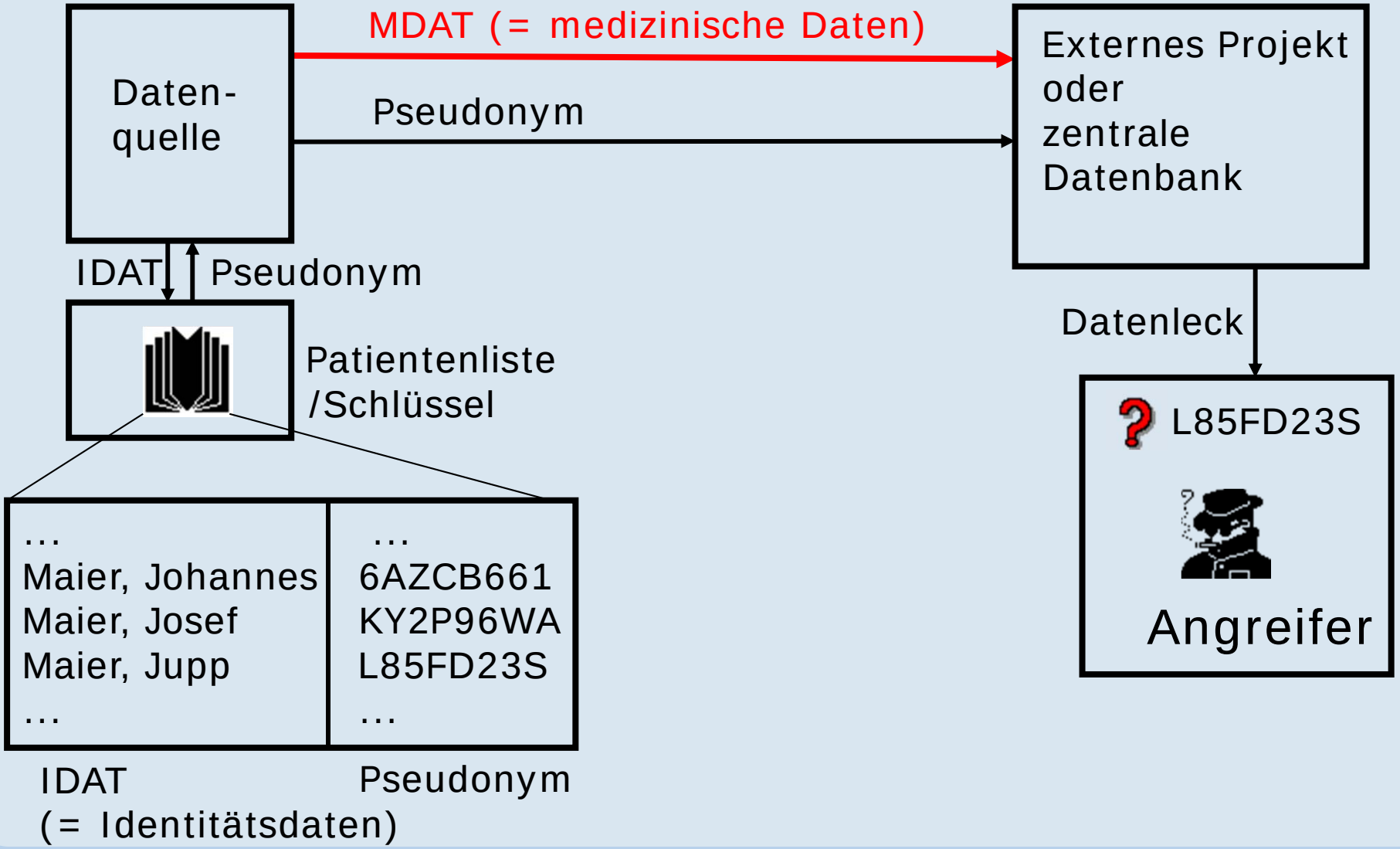
BDSG §3 (6a):

Pseudonymisieren *ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren*.

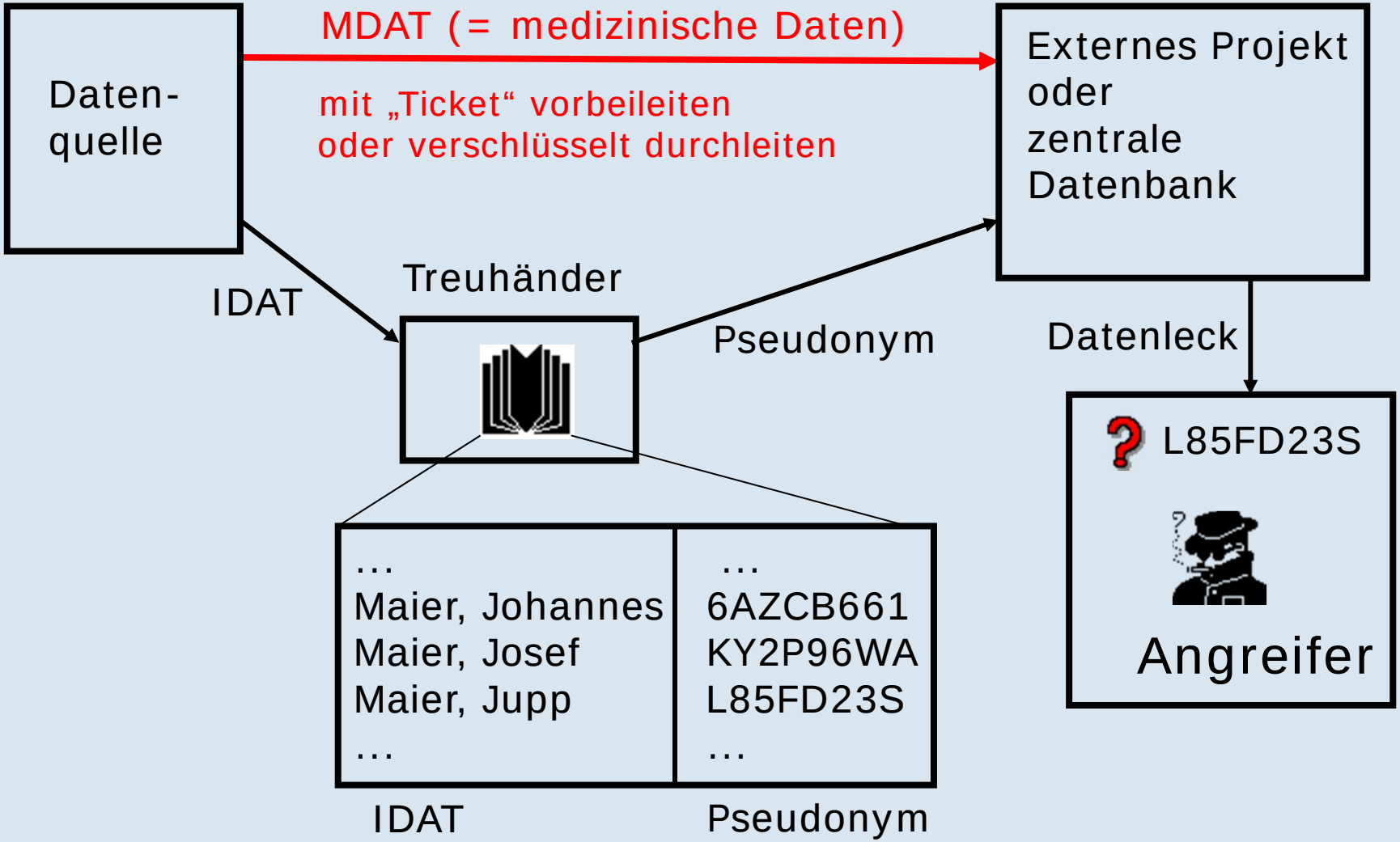
Pseudonymisierung bewahrt die Personenbeziehbarkeit,
schützt sie aber durch einen „Schlüssel“,
der bei einem „Treuhänder“ aufbewahrt wird.

- Pseudonymisierung bedeutet: Ersetzen der identifizierenden Daten durch eine nichtsprechende Zeichenkette.
  - Auch als Codierung/ Verschlüsselung bezeichnet.
  - Schlüssel: Zuordnungsliste oder Schlüssel für kryptographische Transformation,

Pseudonymisierung von Daten: Das Basismodell („Pseudonymisierung an der Quelle“)



Pseudonymisierung von Daten: das Treuhändermodell



Eigenschaften der Pseudonymisierung

- Zusammenführung von Daten aus verschiedenen Quellen möglich
- ... oder von verschiedenen Zeitpunkten.
- Weg zurück zum Patienten für Rückmeldungen offen
- ... oder zur Rekrutierung für neue Studien
- ... oder (bei Biobanken) zum Rückruf von Proben.

Achtung: *Kein* Pseudonym ist

- Initialen + Teile des Geburtsdatums,
- Nummer, die einem größeren Personenkreis bekannt ist (Fall-Nr. im KIS, Versicherungsnummer, ...).

Pseudonymisierung im Vergleich zur Anonymisierung

Pseudonyme Daten *personenbeziehbar* (durch Treuhänder).

- Daher Pseudonymisierung rechtlich *nicht* äquivalent zur Anonymisierung,
  Anonymisierung (fast*) immer erlaubt,
  Pseudonymisierung erfordert Einwilligung.

Aber: Weitergegebene pseudonymisierte Daten gelten als anonymisiert, wenn Empfänger keine Möglichkeit zur Depseudonymisierung hat.

Ständige Beurteilung des RI-Risikos und Einzelfallprüfung bei pseudonymen wie bei anonymen Daten notwendig.

[* s. Proben]

Eignung der Modelle

Basismodell (Pseudonymisierung an der Quelle):

- kleinere Projekte,
- „Umzug“ von Patienten nicht relevant.

Treuhändermodell:

- Zusammenführung von Daten aus verschiedenen Quellen nötig.
- Beobachtung von Patienten über längere Zeiträume nötig.
- Viele Datenquellen mit wenigen Fällen.

Identitätsmanagement

Bei größeren Projekten oft verschiedene Pseudonyme in verschiedenen Bereichen.

Identitätsmanagement

- verwaltet Zuordnung zwischen Pseudonymen und Identitäten
- und zwischen verschiedenen Pseudonymen,
- wirkt bei der Kontaktierung mit.

Z. B. durch Führung einer Patientenliste.