1 Identitätsmanagement mit dem PID-Generator der TMF für das KPOH

KLAUS POMMERENING, MAINZ

Zur Kennzeichnung von Patienten und zur Pseudonymisierung wurde im Kompetenznetz für Pädiatrische Onkologie und Hämatologie (KPOH) ein Werkzeug für die Erzeugung von pseudonymen Patientenidentifikatoren (PID) entwickelt, das durch flexible Konfigurationsmöglichkeiten auch in anderen Forschungsszenarien eingesetzt werden kann. Im generischen Datenschutzkonzept der TMF wird es an zentraler Stelle verwendet.

1.1 Identität und Pseudonym

1.1.1 Daten für Versorgung und Forschung

In medizinischer Versorgung und Forschung werden Daten von Patienten – oder auch gesunden Kontrollpersonen – verarbeitet. Wie man dabei mit der Identität dieser Personen umgehen muss, hängt vom Kontext ab. Es gibt hierfür zwei grundlegend verschiedene Bereiche:

1. den Behandlungskontext („Primärnutzung" von Patientendaten) – hier ist der Patient bekannt und erwartet auch, persönlich und mit Namen angesprochen zu werden. Das künftige Instrument für das Identitätsmanagement in diesem Kontext ist die elektronische Gesundheitskarte (eGK). Ein Master-Patient-Index (MPI) hilft dabei, verteilte Datenbestände richtig zuzuordnen.
2. den Forschungskontext – hierzu zählt jede Art von „Sekundärnutzung" von Patientendaten wie klinische Forschung, medizinische Qualitätssicherung, Versorgungsforschung, epidemiologische Forschung, Registrierung; selbstverständlich können in diesem Bereich auch gezielt Daten erhoben werden, die im Behandlungskontext nicht auftreten.

Zwischen beiden Bereichen hat der Gesetzgeber eine hohe Barriere aufgerichtet: die ärztliche Schweigepflicht. Daten können über diese Barriere hinweg vom Behandlungskontext in den Forschungskontext exportiert werden, wenn eine dieser Voraussetzungen erfüllt ist:

- Eine gesetzliche Regelung schreibt die Datenweitergabe vor; Beispiel: die Meldepflicht im Krebsregistergesetz des Landes Rheinland-Pfalz.
- Es liegt eine schriftliche Einwilligungserklärung des Patienten vor, in der der Umfang der Daten, der Empfängerkreis, die Verwendung und Zweckbestimmung und die vorgesehene Speicherdauer vollständig und abschließend aufgeführt sind.
- Es handelt sich um anonyme Daten, Daten also, die keinen Rückschluss auf die Identität des Betroffenen erlauben. Im Sinne des Bundesdatenschutzgesetzes reicht es schon, wenn ein solcher Rückschluss einen unverhältnismäßigen Aufwand erfordern würde.

In medizinischen Forschungsprojekten ist oft keine dieser Voraussetzungen erfüllt oder erfüllbar. Dennoch kann ein Datenexport unter zusätzlichen Bedingungen und mit zusätzlichen Schutzmaßnahmen möglich sein. Eine der wichtigsten Methoden dafür ist die Pseudonymisierung der Daten.

Pseudonyme dienen dem Identitätsmanagement im medizinischen Forschungskontext.

Pseudonyme werden zu diesem Zweck von vertrauenswürdigen Instanzen – Datentreuhändern oder Trusted Third Parties (TTP) – erzeugt und verwaltet.

In diesem Artikel wird der Begriff „Datentreuhänder" ohne Berücksichtigung seiner rechtlichen Stellung verwendet. Datentreuhänder kann hier, sofern nichts anderes gesagt wird, ein beliebiger, ansonsten von den Handelnden unabhängiger Dritter sein – oder auch ein online ansprechbarer Server, der unter der Hoheit eines unabhängigen Dritten steht und als „elektronischer Datentreuhänder" angesehen werden kann.

1.1.2 Typen von Pseudonymen

Wer in der Informatik von Pseudonymen spricht, denkt zunächst an die von D. CHAUM Anfang der Achtzigerjahre eingeführten „aktiven“ Pseudonyme [Chaum1985]. Diese werden vom Inhaber selbst erzeugt und verwaltet, und der Inhaber muss bei der Nutzung präsent sein. Sie erhalten ihre Gültigkeit und Rechtssicherheit, indem sie mit einem Zertifikat versehen werden, das durch blinde digitale Signatur erzeugt wird. Geeignet sind sie für Rechtsbeziehungen mit Anonymitätsanspruch, z. B. für anonyme Bezahlung von Online-Diensten.

Wir unterscheiden zwei *Grundtypen* von Pseudonymen
1. **aktive Pseudonyme**, die vom Inhaber selbst verwaltet werden,
2. **passive Pseudonyme**, die von einem Datentreuhänder verwaltet werden.

Für die medizinische Forschung ist dieser Typ von Pseudonymen in der Regel nicht brauchbar, da der Patient bei der Nutzung nicht präsent ist, ja überhaupt nicht greifbar sein soll. Daher sind hier „passive“ Pseudonyme nötig; diese werden ohne Mitwirkung des Betroffenen – aber mit seiner Einwilligung – von einem Datentreuhänder erzeugt und verwaltet. Diese Art von Pseudonymen ist technisch sehr viel anspruchsloser und leichter zu verstehen; die erste uns bekannte explizit formulierte Nutzung im Bereich der medizinischen Forschung wurde Anfang der Neunzigerjahre von MICHAELIS und POMMERENING für den Aufbau von Krebsregistern vorgeschlagen [Pommerening1996], fand von da aus Eingang in verschiedene Krebsregistergesetze und führte schließlich zur Aufnahme des Begriffs der Pseudonymisierung ins Bundesdatenschutzgesetz. Solche passiven Pseudonyme sind für den Aufbau von langfristigen Datensammlungen geeignet; man muss allerdings beachten, dass in der Regel – wenn es sich nicht um eine „Einweg-Pseudonymisierung“ handelt – eine Depseudonymisierung mit Hilfe des Datentreuhänders möglich ist. Daher ist *Pseudonymität rechtlich nicht zur Anonymität äquivalent*; insbesondere darf ohne Einwilligung des Betroffenen nicht ohne weiteres pseudonymisiert werden.

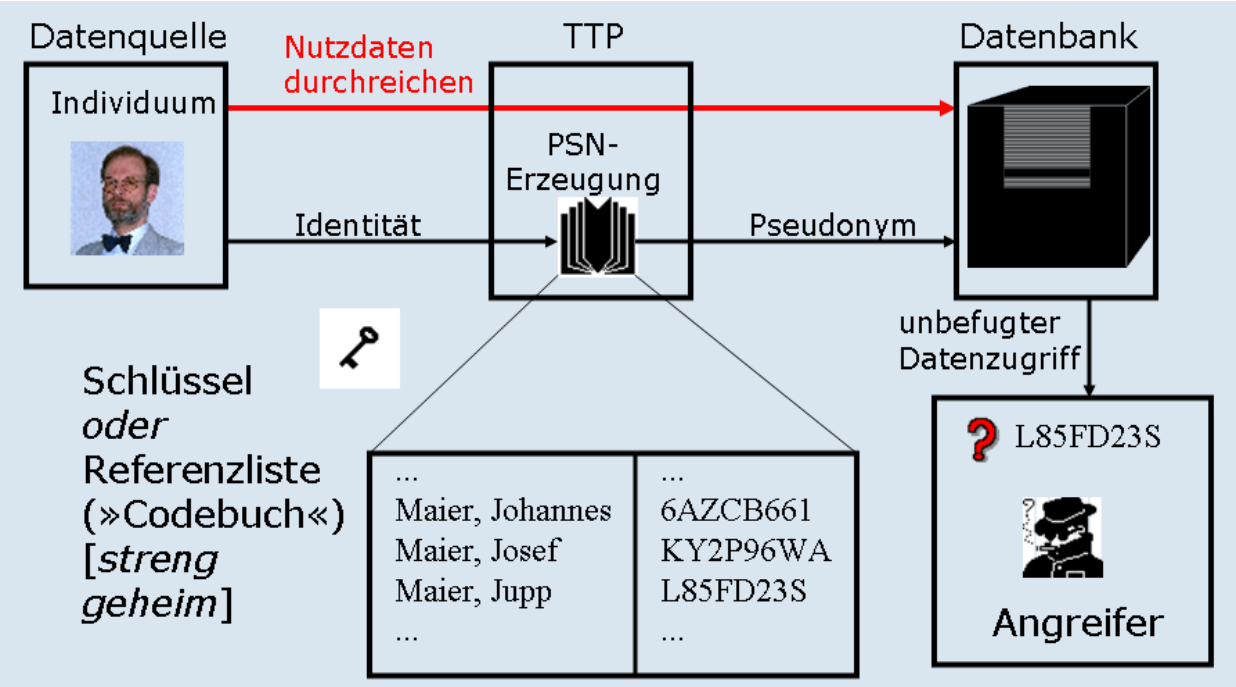

Abb. 1: Passive Pseudonymisierung: das Basismodell

Die Erzeugung von Pseudonymen beim Datentreuhänder kann im Prinzip auf zwei verschiedene Arten geschehen, s. Abb. 1:

1. Der Datentreuhänder ordnet jedem Individuum ein willkürlich gewähltes Pseudonym zu und speichert die Zuordnung in einer Referenzliste („Codebuch-Modell“).
2. Der Datentreuhänder erzeugt aus einem – zuvor standardisierten – Satz von Identitätsdaten (IDAT) ein Pseudonym durch kryptographische Verschlüsselung.

Beide Varianten haben im Kontext der medizinischen Forschung ihren Platz. Einen Spezialfall eines passiven Pseudonyms stellt der vom Arzneimittelgesetz (AMG) geforderte Subject Identification Code (SIC) für klinische Studien dar. Hier ist, abweichend vom Basismodell der Abb.1 das Pseudonym SIC auch der Datenquelle, also dem Prüfarzt bekannt. Daher stellt es keine Abschwächung der Sicherheit dar, wenn

auf einen Datentreuhänder verzichtet und der SIC direkt „an der Quelle“ beim Prüfarzt erzeugt wird, wie es in der Praxis auch oft gehandhabt wird.

1.1.3 Anforderungen an Pseudonyme

Allen Varianten der passiven Pseudonymisierung gemeinsam sind die Anforderungen:

- Der Schlüssel zur Zuordnung – sei es die Referenzliste oder der kryptographische Schlüssel – muss ein Geheimnis des Datentreuhänders (bzw. der Datenquelle) sein.
- Der Datentreuhänder erhält keine weiteren Informationen oder Daten; werden z. B. wie in Abb. 1 die Daten über den Datentreuhänder an eine Datenbank weitergegeben, so werden sämtliche „Nutzdaten“ so verschlüsselt, dass sie erst in der Datenbank wieder entschlüsselt werden können.

Wer immer Zugriff auf die Datenbank erhält, sei es ein Wissenschaftler, der Daten zur Beantwortung wissenschaftlicher Fragestellungen auswerten will, oder ein Unbefugter – ein „Angreifer“ –, kann zwar die Nutzdaten lesen, aber diese nur dem Pseudonym und nicht der eigentlichen betroffenen Person zuordnen. Um eine solche unerwünschte und unerlaubte Herstellung des Personenbezugs – eine „Rückidentifizierung“ oder Re-Identifikation – auszuschließen, muss die Pseudonymerzeugung, egal in welcher Variante, also eine dritte Anforderung erfüllen:

- Das Rückidentifizierungsrisiko muss kontrollierbar sein.

Dies betrifft auch das Risiko, aus den in den Nutzdaten vorhandenen Merkmalen durch Zusatzwissen die Identität einer Person herleiten zu können, z. B. bei seltenen Diagnosen, und gilt natürlich genauso für eine Anonymisierung. Hier ist der Aufwand für einen Angreifer möglichst realistisch abzuschätzen, um dem Anspruch des Bundesdatenschutzgesetzes gerecht zu werden, dass der Aufwand für eine unbefugte Rückidentifizierung unverhältnismäßig sein muss.

Die Qualität der Pseudonymisierung wird durch die Höhe des Rückidentifizierungsrisikos beschrieben.

1.2 Der PID-Generator im KPOH

1.2.1 Ein Patientenidentifikator für die Pädiatrische Onkologie

Die Fachvertreter der Pädiatrischen Onkologie und Hämatologie (POH) in Deutschland arbeiten schon seit langer Zeit eng zusammen. Dabei entstand unter dem Dach der Deutschen Gesellschaft für Pädiatrische Onkologie und Hämatologie (GPOH) eine bundesweit verteilte Kooperationsstruktur, für die seit 1999 durch das Kompetenznetz KPOH auch eine gemeinsame technische Infrastruktur und Vernetzung aufgebaut wurde. Die an dieser Struktur beteiligten organisatorischen Einheiten sind

- die Kliniken, die krebskranke Kinder behandeln,
- etwa 25 multizentrische Therapiestudien für verschiedene Erkrankungen mit den jeweils führenden Fachleuten als Studienleitern,
- Referenzeinrichtungen für Radiologie, Pathologie und Labordiagnostik,
- und schließlich das Deutsche Kinderkrebsregister.

Schon früh entstand für diese Infrastruktur der Wunsch nach einem eindeutigen Patientenidentifikator (PID) für Patienten, die ja nicht überall physisch anwesend sein können, wo ihre Daten verarbeitet werden. Ein solcher wurde als Teil eines KPOH-Projekts entwickelt und Ende 2002 nach einem GPOH-Vorstandsbeschluss eingeführt; insbesondere soll er für neu beginnende Studien verwendet werden. Dieser PID erfüllt die Kriterien:

- Er soll auch als Pseudonym dienen; und zwar soll er für die Therapiestudien die Rolle des SIC im Sinne des Arzneimittelgesetzes übernehmen können.
- Er soll nach einer evtl. nötigen Anonymisierung den Fall im Kinderkrebsregister für die epidemiologische Krebsforschung verfügbar halten.

- Bei der Erzeugung sollen in einem Zwischenschritt „Kontrollnummern“ errechnet und gespeichert werden, die zum pseudonymen Abgleich mit Landeskrebsregistern dienen.

1.2.2 Die Erzeugung von PIDs

Die Software „PID-Generator“ wurde am Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI) der Universität Mainz unter der Projektleitung des Autors entwickelt; die Hauptentwickler und -betreuer waren Markus Wagner (2000 bis 2003), Jutta Glock geb. Moormann (2003 bis 2005), Murat Sariyar (seit 2005). Zur Verwendung durch die GPOH wurde am IMBEI ein zentraler Service aufgesetzt, über den auf verschiedene Weise PIDs bezogen werden können:

- Für eine interaktive Erzeugung wird ein Web-Formular angeboten, s. Abb. 2; der ausgegebene PID kann per „Copy & Paste“ in andere Erfassungsmasken übernommen werden.
- Für größere Datenmengen ist ein Batch-Betrieb, angestoßen durch den Server-Administrator möglich; auf diese Weise wurden z. B. über 44000 „Altfälle“ des Kinderkrebsregisters nachträglich mit einem PID versehen.
- Über eine SOAP-Schnittstelle ist der Web-Service auch direkt in andere Datenerfassungsprogramme einbindbar, z. B. in Studiensoftware.

Insgesamt wurden für die Pädiatrische Onkologie bis Ende 2008 über 57000 PIDs erzeugt.

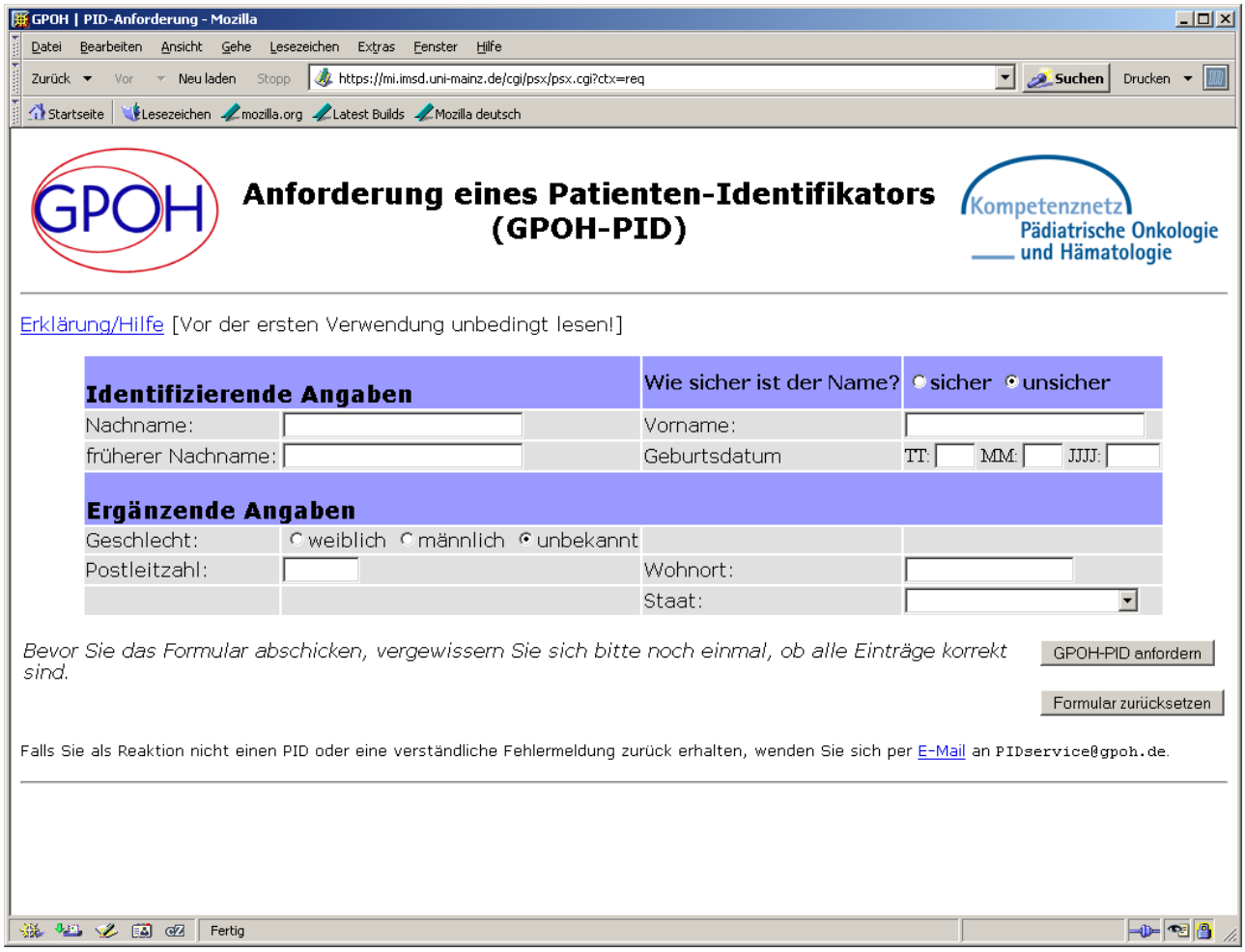

Abb. 2: Bildschirmmaske zur Anforderung eines PID

Den Ablauf der PID-Erzeugung bzw. des PID-Abrufs zeigt Abb. 3. Den algorithmischen Kern des PID-Generators bilden die beiden Komponenten „Match-Algorithmus“ und „Erzeugungs-Algorithmus“.

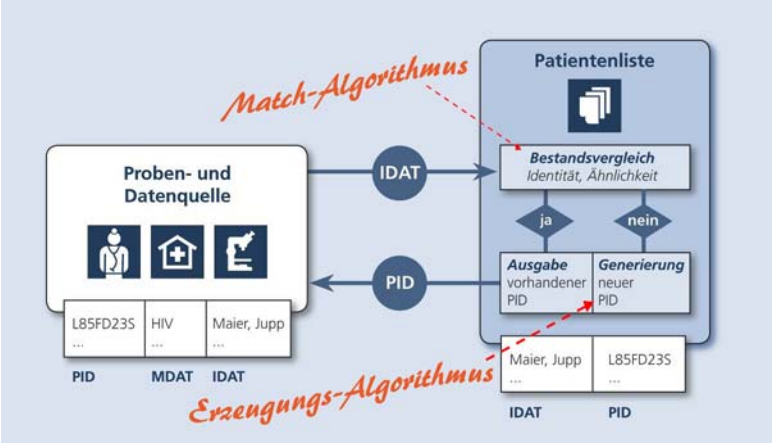

Abb. 3: Ablauf und Algorithmen der PID-Erzeugung

1.2.3 Der Match-Algorithmus

Der Match-Algorithmus soll durch Bestandsvergleich feststellen, ob ein neu eingegebener Fall schon in der Datenbank des PID-Generators vorhanden ist. Hintergrund ist die Erfahrung, dass Identitätsdaten oft fehlerhaft erfasst werden, wodurch

- Synonymfehler – d. h., zur selben Person werden mehrere Datensätze angelegt –
- und Homonymfehler – d. h., verschiedene Personen werden dem gleichen Datensatz zugeordnet –

entstehen. In einem medizinischen Behandlungskontext – der ja bei klinischen Studien durchaus vorhanden ist – sind insbesondere Homonymfehler völlig untragbar. Im reinen Forschungskontext werden je nach Fragestellung Fehlerraten von wenigen Prozent toleriert, sind aber natürlich auch möglichst gering zu halten. Wie auch immer – nach einer Pseudonymisierung ist eine Korrektur nicht mehr möglich. Daher ist das bestmögliche Abfangen dieser Fehler als Vorbereitung der Pseudonymisierung ein kritischer Teil der Datenqualitäts-Sicherung und daher stets integraler Bestandteil des pseudonymen Identitäts-Managements.

> Die Qualität der Daten muss vor der Pseudonymisierung gesichert werden.

Für die Pädiatrische Onkologie wird bisher ein deterministischer Abgleich-Algorithmus eingesetzt, der auch Namensänderungen und verschiedene Schreibvarianten eines Namens sowie phonetische Ähnlichkeit berücksichtigt.

Als erste Stufe der Evaluation [Glock2006] wurden Funktionstests mit halbfiktiven und realen Daten durchgeführt. Dabei wurden keine Homonymfehler in ca 44000 Datensätzen entdeckt, aber 22 Synonyme. Damit ist der für den PID-Generator der GPOH gewählte Datensatz, s. Abb. 2, gut geeignet. In einer zweiten Stufe [Sariyar2009] wurden verschiedene Match-Verfahren, u. a. stochastische Verfahren, Entscheidungsbäume und Support Vector Machines, vergleichend evaluiert, mit und ohne phonetische Vergleiche. Dabei stellte sich heraus, dass im Kontext des PID-Generators das „klassische“ Record Linkage nach Newcombe [Newcombe1988] den anderen Verfahren leicht bis deutlich überlegen ist, zumindest bei geschickter Aufbereitung.

Als Fazit der Evaluation kann man festhalten:

> Homonym- und vor allem Synonymfehler sind beim Datenabgleich nie ganz auszuschließen. Eine gute Phonetik reduziert die Synonymfehler.

1.2.4 Der Erzeugungs-Algorithmus

Der PID wird im PID-Generator nicht direkt aus den Identitätsdaten erzeugt, sondern als kryptographisch verschlüsselte laufende Nummer. Das bedeutet, dass in der Datenbank die Zuordnungsliste mitgeführt werden muss; in ihr können die Identitätsdaten als Klartext oder Einweg-verschlüsselt – als „Kontrollnummern“ – abgelegt werden. Der PID-Algorithmus hängt also von einem Schlüssel ab, der in jedem Anwendungsumfeld gesondert gesetzt werden sollte, um die Entstehung eines kontext-übergreifenden Personenkennzeichens zu verhindern.

Für die Gestalt der PIDs gibt es zwei Optionen. Müssen sie nur maschinenlesbar sein, kann direkt die aus der Verschlüsselung entstandene Bitkette, also das AES-Chiffrat der laufenden Nummer, verwendet werden. Sollen sie auch für menschliche Augen lesbar sein, sollten sie aus leicht erfassbaren und schwer zu verwechselnden Zeichen bestehen und auch gegen fehlerhafte Übertragung resistent sein. Zu diesem Zweck wurde für den PID-Generator ein spezieller fehlererkennender und -korrigierender mathematischer Code entwickelt [Faldum2005], der nach seinem Erfinder FALDUM-Code heißen soll. Für die Darstellung werden 8 Zeichen verwendet, wovon die letzten beiden Prüfzeichen sind. Als Zeichensatz dienen die Großbuchstaben und Ziffern, wobei die Buchstaben B, I, O, S wegen der Verwechslungsmöglichkeit mit den Ziffern 8, 1, 0, 5 ausgeschlossen werden. Es gibt also $32 = 2^5$ nutzbare Zeichen und somit $(2^5)^6 = 2^{30}$ nutzbare Codewörter, also gut eine Milliarde. Die Prüfzeichen erlauben, bis zu zwei Fehler zu entdecken oder einen Fehler zu korrigieren; als Besonderheit speziell dieses Codes kann auch eine Vertauschung zweier Nachbarzeichen korrigiert werden.

Die Erzeugung menschenlesbarer PIDs ist gegenüber den maschinenlesbaren wegen des kleineren Wertevorrats notwendig kryptographisch weniger sicher. Da die Verschlüsselung direkt aber nur die laufende Nummer schützt, die selbst kaum Information enthält, wirkt sich diese kryptographische Schwäche praktisch nicht auf die Sicherheit des Gesamtverfahrens aus.

Im laufenden Betrieb darf der PID-Algorithmus samt seinem Schlüssel *nicht mehr* geändert werden, da dann alle bisher erzeugten PIDs ungültig würden. Aber bei einer Kompromittierung des Verfahrens kommt ohnehin nur ein völliges Verwerfen aller bisherigen PIDs und Neuerzeugung in Frage.

Der Match-Algorithmus darf im laufenden Betrieb jederzeit geändert werden, der Erzeugungs-Algorithmus nie.

1.2.5 Der PID-Generator als Werkzeug

Der PID-Generator steht als Software auch für andere Projekte zur Verfügung. Produkt-Information, Manual und andere Informationen sind online unter http://www.staff.uni-mainz.de/pommeren/PID/ erhältlich, der Quellcode selbst über die TMF.

Der PID-Generator dient dazu, eine definierte Population mit pseudonymen (nichtsprechenden) Identifikatoren zu versehen. Dazu wird jeweils ein Satz von personenidentifizierenden Daten in einen PID transformiert. Berücksichtigt wird dabei, dass die personenidentifizierenden Merkmale fehlerbehaftet und zeitlich veränderlich sein können. Der Kern des PID-Generators ist eine Datenbank, die die bisher eingegebenen Fälle – die personenidentifizierenden Daten im Klartext oder in umkehrbar oder unumkehrbar verschlüsselter Form – zusammen mit dem jeweiligen PID speichert.

Der PID-Generator ist als Web-Service konzipiert. Dazu wird das Programm über die CGI-Schnittstelle eines Webservers angesprochen; als Benutzungsoberfläche dienen konfigurierbare HTML-Seiten. Darüber hinaus ist ein interaktiver Konsolenbetrieb sowie ein Batch-Betrieb möglich. Im letzteren Fall ist die Ergebnis-Rückmeldung in eine Datei möglich. Für administrative Datenpflege, etwa bei nachträglichem Erkennen eines Homonyms, ist ein Direktzugriff auf die Datenbank zu nutzen. Eine SOAP-Schnittstelle zur Einbindung in bestehende RDE-Systeme ist als externer Modul vorhanden.

Die Flexibilität des PID-Generators wird über seine Konfigurationsdatei zugänglich gemacht. In dieser können u. a. definiert werden:

- die Struktur der Eingabedatensätze (Feldnamen, Typ, Feldgrenzen, Transformationsoptionen, Abgleichsoptionen),
- der Ablauf des Match-Verfahrens als Entscheidungsbaum einschließlich der Update-Strategie für Datenbank-Einträge,
- die Datenbank-Verbindung,
- die Log-Datei und der Umfang der Log-Aktivitäten,
- die Nutzung eines Verzeichnisdienstes über LDAP,
- Templates für die HTML-Seiten, die die Benutzungsoberfläche ausmachen,
- die Meldungstexte für die verschiedenen Ausgänge des Match-Verfahrens,
- die Schlüssel für die verschiedenen kryptographischen Operationen.

Das Datenbankschema wird aus der Konfigurationsdatei erzeugt. Die Transformationsoptionen umfassen unter anderem die Normalisierung von Namen, die Erzeugung phonetischer Codes und kryptographische Verfahren. Die Verwendung der Krankenversicherten-Nummer als identifizierendes Merkmal ist im PID-Generator grundsätzlich vorgesehen, für die Pädiatrische Onkologie aber nicht sinnvoll umsetzbar.

Der Ablauf des Match-Verfahrens kann durch die Attribute „sicher“ und „unsicher“ für die zu vergleichenden Datensätze beeinflusst werden. Eine besondere Rolle spielen die Krankenversicherten-Nummern, die, falls vorhanden, bei exakter Übereinstimmung stets zu einem Match führen. Eine Umkonfiguration des Match-Verfahrens im laufenden Betrieb ist jederzeit möglich; insbesondere können Homonym- und Synonymfehlerraten nach jeweiligem Bedarf austariert werden. Auch der Einbau anderer Match-Algorithmen ist im laufenden Betrieb jederzeit möglich, erfordert aber eine Neukompilierung der Software.

Der PID-Generator ist ein flexibles Werkzeug, das sich für verschiedene Szenarien und Anforderungen passend konfigurieren lässt.

Der PID-Generator ist in reinem C entwickelt und sollte sich auf jedem UNIX-System problemlos kompilieren lassen; getestet wurde das unter verschiedenen Linux-Versionen und OpenBSD. Für MS-Windows-Systeme ist eine installationsfertige Version vorhanden. Als Datenbank wird PostgreSQL direkt unterstützt; andere Datenbanken können über eine ODBC-Schnittstelle angebunden werden.

Mehrere freie Software-Pakete wurden als Fremdsoftware eingebaut: Hash-Algorithmen, das Verschlüsselungsverfahren AES, die Kölner Phonetik [Postel1969], Hannoveraner Phonetik [Michael1999]. Einige der Verfahren wie die Normalisierung von Namen und die Erzeugung von Kontrollnummern sind Weiterentwicklungen von Algorithmen des Landeskrebsregisters Rheinland-Pfalz.

1.3 Der PID-Generator im TMF-Datenschutzkonzept

1.3.1 Einsatzszenarien

Im generischen Datenschutzkonzept der TMF [Reng2006] werden zwei unterschiedlichen Szenarien beschrieben; bei beiden wird der PID-Genrator, allerdings in leicht unterschiedlicher Rolle, eingesetzt:

- (Modell A) Der PID dient als pseudonymes Kennzeichen in einer einrichtungsübergreifenden Versorgungsdatenbank, zu der parallel eine separate Patientenliste geführt wird. Hier ist die Speicherung pseudonym, aber für Berechtigte ist ein personenbezogener Online-Zugriff möglich, s. Abb. 4.
- (Modell B) Der PID wird als zusätzliches Kennzeichen zu den Identitätsdaten (IDAT) behandelt. Für eine „Forschungsdatenbank“ wird ein Pseudonym durch kryptographische Verschlüsselung des PID hergestellt; eine TTP dient hierfür als Pseudonymisierungsdienst. In diesem Modell sind Speicherung und Zugriff nur pseudonym. Zur Erhöhung der Datenqualität vor der Pseudonymisierung ist eine fehlertolerante PID-Erzeugung nötig, s. Abb. 5.

Die beiden Modelle haben die pseudonyme Speicherung gemeinsam. Sie unterscheiden sich darin, ob ein personenbezogener oder nur ein pseudonymer Zugriff auf medizinische Daten möglich ist.

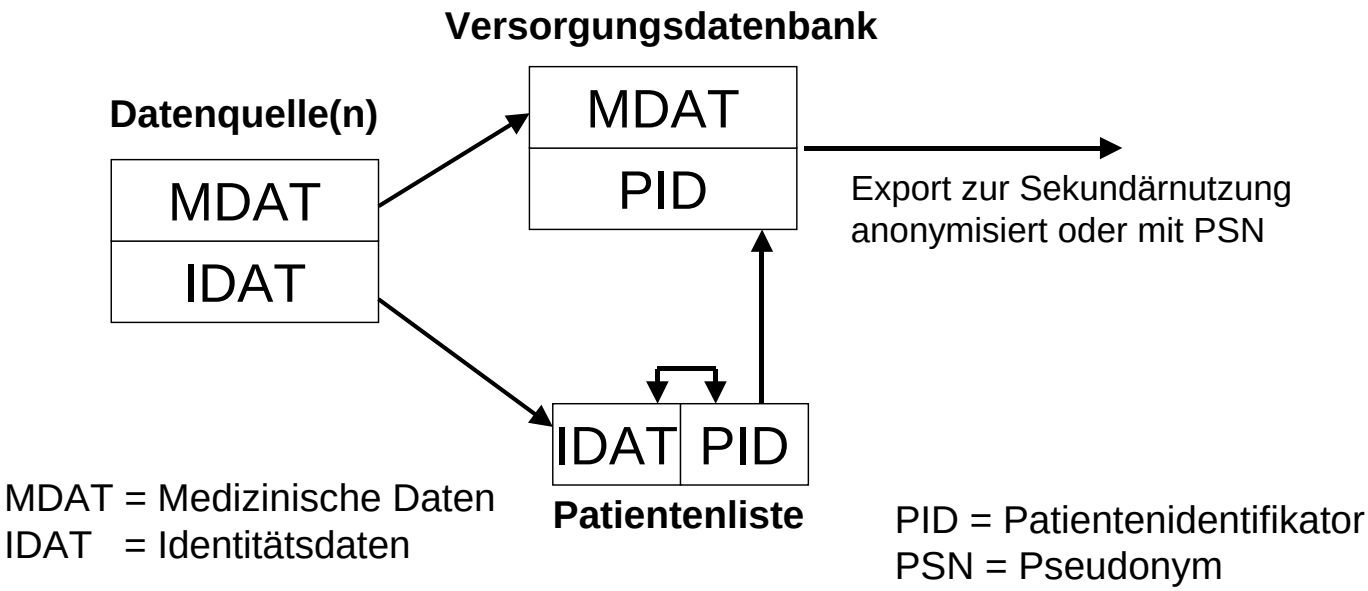

Abb. 4: Identitätsmanagement im TMF-Modell A (patientennahe Versorgungsdatenbank)

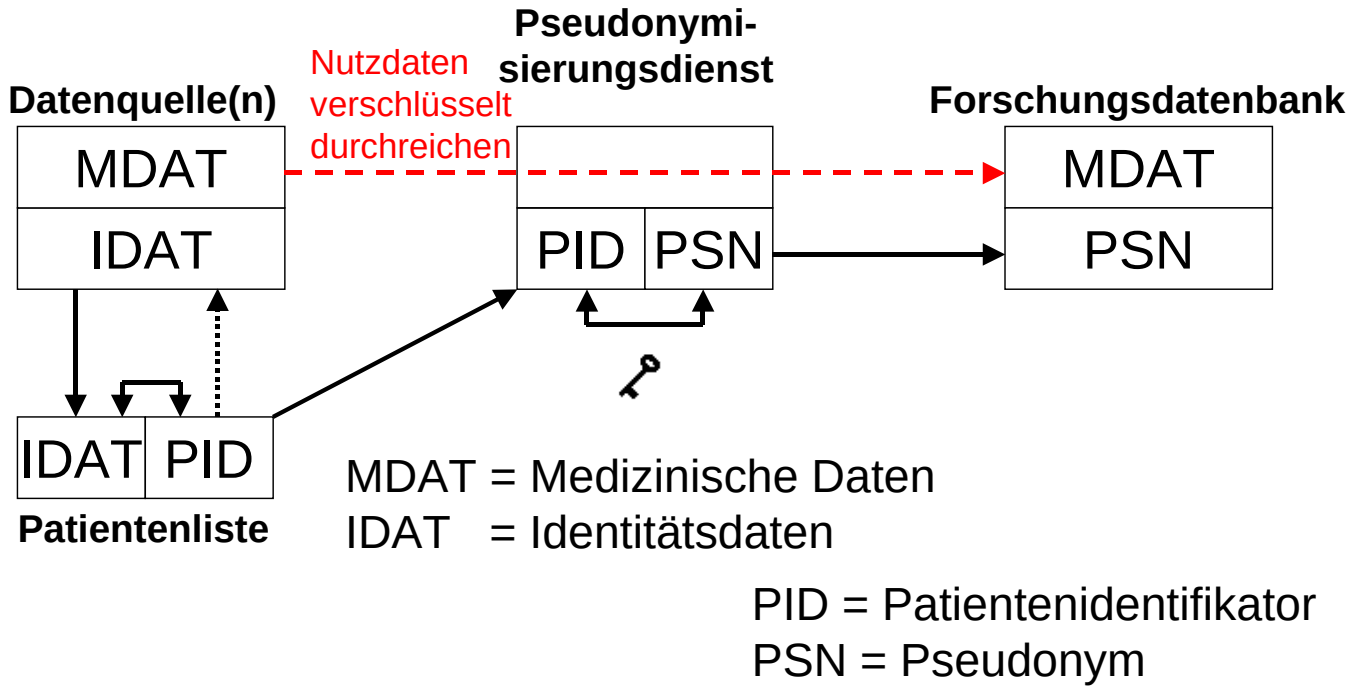

Abb. 5: Identitätsmanagement im TMF-Modell B (patientenferne Forschungsdatenbank)

1.3.2 Die Revision des Datenschutzkonzepts

In der zur Zeit stattfindenden Überarbeitung des generischen Datenschutzkonzepts wird eine einheitliche umfassende Systemarchitektur diese beiden Modelle verallgemeinern, s. Abb. 6. Diese Architektur ist modular: Es werden bis zu fünf Arten von Datenbanken unterschieden, die alle in einem Netz, auch mehrfach, vorkommen können:

- Versorgungsdatenbank,
- Studiendatenbank,
- Forschungsdatenbank,
- Bilddatenbank,
- Biomaterialbank.

Diese unterscheiden sich nach Rahmenanforderungen, Speicherart und Zugriffsart und verwenden jeweils ihre eigenen Pseudonyme. Zentral zwischen allen steht eine, möglicherweise auch verteilte, Identitätsmanagement-Komponente, die bei Bedarf den Übergang zwischen diesen verschiedenen Pseudonymen und auch der wahren Identität herstellt. Die Architektur ist skalierbar: Je nach Verhältnismäßigkeit sind verschiedene Vereinfachungen möglich.

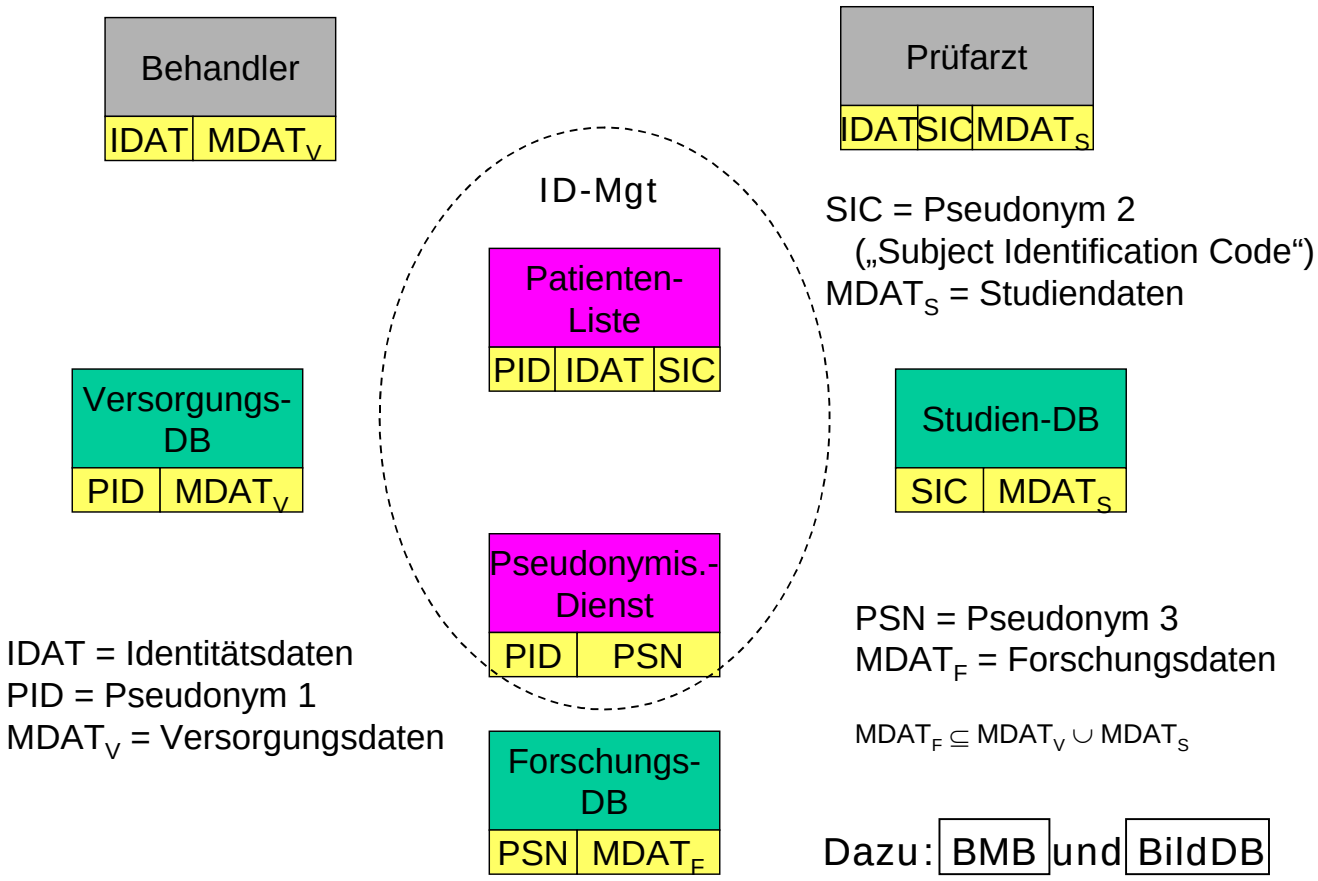

Abb. 6: Referenzmodell im künftigen revidierten Datenschutzkonzept der TMF: Datenbanken und Daten

Die Versorgungs-Datenbank (VDB) enthält Daten, die für die Versorgung des Patienten relevant sind; sie steht im unmittelbaren Behandlungskontext, ist aber einrichtungsübergreifend und wird daher pseudonym (PID) geführt. Behandler haben einen personenbezogenen Zugriff auf die Daten ihrer Patienten (lesend und schreibend). Der Zugriff geschieht mit Hilfe der Patientenliste (ID-Management für Patienten) und mit Hilfe eines Verzeichnisdienstes (ID-Management für Benutzer). Dieser Modul ist eine Weiterentwicklung aus dem TMF-Modell A. Er ist bei jeder Art von einrichtungsübergreifender Versorgung relevant; im Kontext der medizinischen Forschung wird er dort benötigt, wo langfristige Beobachtungsstudien durchgeführt werden, etwa bei seltenen oder chronischen Erkrankungen.

Die Studien-Datenbank (SDB) dient zur Durchführung klinischer Studien nach den Regularien des AMG und der guten klinischen Praxis (GCP). Sie enthält Daten zum Patienten, die für die Studie relevant sind. Die Überschneidung mit den Daten der reinen Versorgungsdokumentation ist groß. Die SDB steht im unmittelbaren Behandlungskontext, soweit es um Zugriffe durch den Prüfarzt geht; sie steht im Forschungskontext, wenn es um Zugriffe durch den „Sponsor“ oder Studienleiter geht. Sie ist, zumindest bei multizentrischen Studien, einrichtungsübergreifend. Sie wird konform zum AMG pseudonym (SIC) geführt. Prüfärzte haben einen personenbezogenen Zugriff auf die Daten ihrer Patienten (lesend und schreibend) und kennen den SIC. Dieses Szenario entspricht dem Modell der Pädiatrischen Onkologie, wobei der GPOH-PID die Rolle des SIC einnimmt.

Die Forschungs-Datenbank (FDB) dient zur Langzeitspeicherung pseudonymisierter medizinischer Daten für spätere Forschungsprojekte – direkt zur epidemiologischen Forschung, indirekt zur Rekrutierung geeigneter Fälle für neue klinische oder epidemiologische Forschung. Sie bietet den nochmals pseudonymisierten Export geeigneter Daten und stellt eine Weiterentwicklung des TMF-Modells B dar.

Das Identitätsmanagement ist eine zentrale Funktion eines medizinischen Forschungsnetzes. Seine Werkzeuge sind PID-Generator und Pseudonymisierungsdienst.

1.3.3 Weiterentwicklung der Werkzeuge

Aus der Revision des generischen Datenschutzkonzepts ergeben sich Anforderungen an die Weiterentwicklung der Werkzeuge des Identitätsmanagements, PID-Generator und Pseudonymisierungsdienst:

- Das Identitätsmanagement ist um weitere pseudonyme Kennzeichen zu erweitern; diese können durch kryptographische Verschlüsselung des PID gewonnen werden. Für Studiendatenbanken wird jeweils ein SIC, für Bilddatenbanken ein BildID, für Proben und genetische Analysen im Rahmen einer Biomaterialbank ein LabID benötigt.
- Für den Match-Algorithmus im PID-Generator sind weitere Wahlmöglichkeiten zu schaffen.
- Die Internationalisierung des PID-Generators ist durch Zulassung verschiedener Zeichensätze und Transkriptionsregeln, die in Unicode umgesetzt werden, sowie durch Phonetik in verschiedenen relevanten Sprachen zu unterstützen.

Die rechtlichen und technischen Anforderungen an die nötigen Datentreuhänderdienste werden in Projekten der TMF geklärt; die Publikation der Ergebnisberichte ist im Druck.

Literatur

[Chaum1985] Chaum, D.: Security without identification: Transaction systems to make Big Brother obsolete. Communications of the ACM 28 (1985), 1030-1044.

[Faldum2005] Faldum, A., Pommerening, K.: An optimal code for patient identifiers. Computer Methods and Programs in Biomedicine 79 (2005), 81-88; online unter http://else.hebis.de/cgi-bin/sciserv.pl?collection=journals&journal=01692607&issue=v79i0001&article=81_aocfpi

[Glock2006] Glock, J., Herold R., Pommerening K.: Personal identifiers in medical research networks: Evaluation of the personal identifier generator in the Competence Network Paediatric Oncology and Haematology. GMD Med Inform Biom Epidemiol. 2/2 (2006), Doc06; online unter http://www.egms.de/en/journals/mibe/2006-2/mibe000025.shtml

[Michael1999] Michael J. Doppelgänger gesucht – Ein Programm für kontextsensitive phonetische Textumwandlung. c't. 25/1999; 252-61.

[Newcombe1988] Newcombe, H.B.: Handbook of Record Linkage. New York: Oxford University Press 1988

[Pommerening1996] Pommerening, K., Miller, M., Schmidtmann, I., Michaelis J.: Pseudonyms for cancer registry. Methods of Information in Medicine 35 (1996), 112-121.

[Postel1969] Postel, H.J.: Die Kölner Phonetik – Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM-Nachrichten. 19 (1969); 925-31.

[Reng2006] Reng, C.-M., Debold, P., Specker, C., Pommerening K.: Generische Lösungen der TMF zum Datenschutz für die Forschungsnetze der Medizin. Medizinisch Wissenschaftliche Verlagsgesellschaft, München 2006.

[Sariyar2009] Sariyar, M.: Record Linkage im Kontext von iterativen Eingaben. Dissertation, Johannes-Gutenberg-Universität Mainz, 2009.