Massenpseudonymisierung von persönlichen medizinischen Daten

Johannes Heurix[1] · Thomas Neubauer[2]

[1]Technische Universität Wien
heurix@ifs.tuwien.ac.at

[2]Secure Business Austria
neubauer@securityresearch.ac.at

Zusammenfassung

Im Gesundheitswesen spielt die ständige Verfügbarkeit personenbezogener Daten eine zentrale Rolle. Durch die Digitalisierung von medizinischen Daten und deren Verwaltung in elektronischen Gesundheitsakten (ELGA), werden täglich große Mengen an Gesundheitsdaten erstellt und verwaltet. Diese Informationen müssen aus gesetzlichen Gründen für eine Dauer von bis zu 30 Jahren gespeichert werden, haben zugleich allerdings keinen Nutzen für Forschungseinrichtungen, da gesetzliche Regelungen die Verwendung von Dokumenten, die sowohl persönliche als auch medizinische Daten enthalten, untersagen, was zu kostspieligen Datenbeschaffungsphasen und unvollständigen Rohdatensammlungen führt. Dieser Beitrag präsentiert eine neue Methode zur Massenpseudonymisierung medizinischer Daten mit dem Ziel, klinische Studien mit bereits existierenden medizinischen Daten zu versorgen, wobei durch die Pseudonymisierung der entsprechenden persönlichen Daten der Schutz der Privatsphäre der beteiligten Patienten gewährleistet wird. Dieser Ansatz behebt eines der Hauptprobleme in der klinischen Forschung, indem Forschungseinrichtungen ein System zur Pseudonymisierung großer Datenmengen, die zur effizienten und effektiven Durchführung von Studien benötigt werden, zur Verfügung gestellt wird.

1 Einführung

Die elektronische Gesundheitsakte (ELGA) verspricht, die Kommunikation zwischen den Dienstleistern im Gesundheitswesen und somit die Qualität der Patientenbetreuung zu verbessern, sowie Kosten zu senken. Tatsächlich reduziert die Anwendung der ELGA das Auftreten von unerwünschten Nebenwirkungen von Medikamenten (die für die Kosten von etwa 175 Milliarden Dollar und für mehr als 200.000 Todesfälle pro Jahr in den USA [EG01] verantwortlich sind), da sie die Ärzte und ihre Mitarbeiter mit Richtlinien hinsichtlich dem Zusammenwirken von Medikamenten versorgen und in deren Entscheidungsfindung unterstützen. Darüber hinaus verspricht die ELGA massive Einsparungen bei der Verwaltung medizinischer Daten. Allerdings führt die Sekundärdatenanalyse elektronisch gespeicherter Patientendaten zu Problemen in Bezug auf den Datenschutz. Tatsächlich ist die Diskussion des Datenschutzes im heutigen Gesundheitswesen eines der zentralen Themen und erfordert einen Kompromiss zwischen den datenschutzrechtlichen Bedürfnissen der Patienten auf der einen Seite und der Forderung nach gesteigerter Effizienz im Gesundheitssystem und der besseren Verfügbarkeit

P. Horster (Hrsg.) · D•A•CH Security 2009 · syssec (2009) 1-12.

der Daten für klinische Studien auf der anderen Seite. Eine Offenlegung sensibler persönlicher medizinischer Daten kann ungünstige Auswirkungen für den Patienten wie z.B. verweigerte Versicherungen zur Folge haben. Um den Missbrauch von Patientendaten zu verhindern, wurden Gesetze wie der US-amerikanische Health Insurance Portability and Accountability Act (HIPAA) [Uni06] oder die europäische Direktive 95/46/EC [Eur95] eingeführt. Für den Schutz der Privatsphäre von Patienten wurden zahlreiche Methoden vorgeschlagen (vgl. [FH01]). Bestehende Ansätze jedoch (i) richten sich nicht nach aktuellen gesetzlichen Anforderungen (vgl. [Eur95, Hin03, HGG05, U.S03, U.S96]), (ii) erfüllen oft nicht die grundlegenden Sicherheitsanforderungen (vgl. [SAG+05, RNG+07, BC96]), oder (iii) sind nicht für klinische Studien anwendbar. Dieser Beitrag widmet sich den Schwächen der bestehenden Ansätze und schlägt eine Erweiterung des PIPE-Systems (vgl. [RNG+07]) für Massenpseudonymisierung vor. In PIPE spielt der Patient als Dateneigentümer seiner medizinischen Daten eine zentrale Rolle, wobei er über seine Daten volle Kontrolle ausübt und Vertrauenspersonen autorisieren kann, auf diese zugreifen zu können. Analog zu Massensignaturen [HK03], wo große Mengen an Dokumenten mit nur minimalem aktivem menschlichem Eingreifen automatisch digital unterschrieben werden, zielt das Konzept der Massenpseudonymisierung darauf ab, große Mengen an Patientenunterlagen automatisiert zu pseudonymisieren, wobei eine Gruppe von so genannten Datenadministratoren als (Interims-) Dateneigentümer agiert. Dieser Lösungsansatz schließt eine wesentliche Lücke in der klinischen Forschung und stellt Forschungseinrichtungen ein sicheres System für eine effektivere und effizientere Ausführung klinischer Studien zur Verfügung, indem unter Einhaltung des Datenschutzes existierendes medizinisches Datenmaterial genutzt werden kann.

2 Pseudonymisierung zur Verringerung datenschutzrechtlicher Bedenken

Anonymisierung und Verschlüsselung sind zwei anerkannte Techniken, um die Privatsphäre eines Patienten zu schützen: Anonymisierung bedeutet, den Patienten identifizierende Informationen aus den medizinischen Unterlagen zu entfernen, sodass die Verbindung zwischen einem Patienten und seinen medizinischen Unterlagen nicht mehr hergestellt werden kann. Der Nachteil dieser Methode ist die Unumkehrbarkeit, sodass sie nur für die Sekundärnutzung von medizinischen Daten verwendet werden kann. Die Verschlüsselung medizinischer Daten hingegen sichert die Privatsphäre der Patienten und die Vertraulichkeit der Daten, indem diese nur für autorisierte Personen zugänglich gemacht werden, wobei die Verschlüsselung zwar umkehrbar ist, jedoch die anonyme Nutzung medizinischer Daten für klinische Studien unmöglich wird, da der Patient durch die erforderliche ausdrückliche Genehmigung (Entschlüsselung mit seinem geheimen Schlüssel) seine Identität preisgeben muss. Pseudonymisierung ist eine Technik, bei der identifizierende Daten durch einen speziellen Bezeichner ersetzt werden, der ohne das Wissen um ein bestimmtes Geheimnis nicht mit den Nutzdaten assoziiert werden kann [PK05, RNG+07, Tai04]. Pseudonymisierung erlaubt eine Assoziierung der Nutzdaten mit einem bestimmten Patienten nur unter spezifischen und kontrollierten Rahmenbedingungen und überwindet die Schwächen sowohl der Anonymisierung als auch der Verschlüsselung, indem im Wesentlichen die medizinischen Daten anonymisiert gespeichert werden, um deren Verwendung in klinischen Studien zu ermöglichen, jedoch die Umkehrbarkeit dieses Prozesses trotzdem sichergestellt wird, indem bei Bedarf die Verbindung zum Patienten wieder hergestellt werden kann. Allerdings haben bestehende Ansätze und Systeme verschiedenste Mängel. Das von Thielscher et al. (vgl. [TGU+05]) entwickelte Verfahren stützt sich auf eine zentral

gespeicherte Patient-Pseudonym-Liste als Sicherungsmechanismus für den Fall eines Verlustes der Smart Card des Patienten. Um die Liste gegen Angriffe zu schützen, wird diese offline betrieben, was ein höheres Maß an Sicherheit gewährleistet, bis eine Person innerhalb des Systems das Ziel eines erfolgreichen Social Engineering-Angriffs wird [Mar05, Tho04] oder ein Angreifer physischen Zugriff auf den Computer mit der Liste erlangt. Die von Pommerening (vgl. [Pom94, PR04]) entwickelten Ansätze verwenden eine Kombination aus Hash-Techniken und Verschlüsselung, wobei ein einzelner zentraler Geheimschlüssel verwendet wird und dieser daher eine Schwachstelle darstellt, da ein Angreifer, der diesen einzelnen Schlüssel kennt, Zugang zu allen medizinischen Patientendaten erhält. Der von Peterson (vgl. [Pet03]) entwickelte Ansatz hat mehrere Nachteile: Da alle Schlüssel, die zur Dechiffrierung der medizinischen Daten nötig sind, in der Datenbank gespeichert sind, kann ein Angreifer, der Zugang auf diese Datenbank erlangt, alle Informationen entschlüsseln. Noch gravierender ist, dass neben den Schlüsseln auch das Passwort in derselben Datenbank gespeichert ist, womit ein Angreifer die dort vorhandenen Daten verändern kann. Die von Schmidt et al. (vgl. [SSP+01]) und dem vom deutschen Bundesministerium für Gesundheit geförderten Fraunhofer Institut (vgl. [Fra05, Cau06]) vorgeschlagenen Architekturen stützen sich auf eine vollständige Verschlüsselung der Daten, was für klinische Studien nicht praktikabel ist.

3 Systemarchitektur

Das PIPE-System ist als mehrstufiges Hüllenmodell mit drei verschiedenen Schichten realisiert (Abbildung 1). Jede Schicht ist für eine Stufe des Datenzugriffsprozesses verantwortlich. Der User muss alle Schichten passieren, um die eigentlichen Patientenunterlagen abrufen zu können. Die äußere Hülle, die Authentifizierungsschicht, wird durch das äußere, asymmetrische

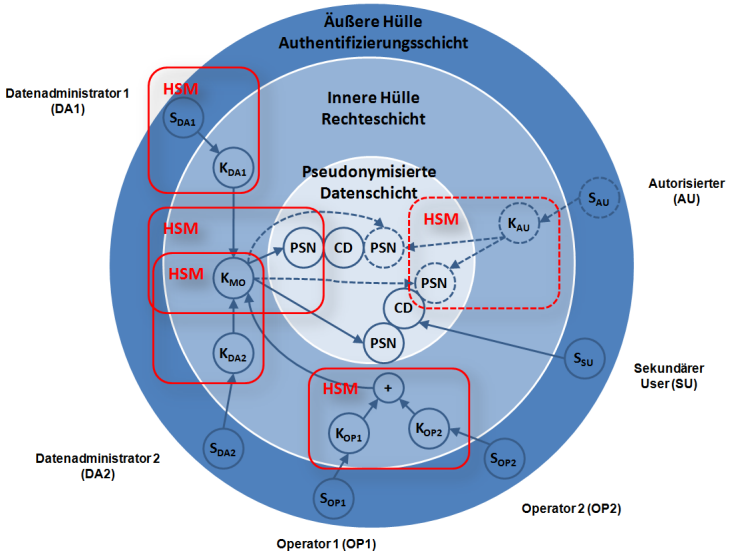

Abb. 1: Modifizierte PIPE-Hüllenarchitektur für die Massenpseudonymisierung

Schlüsselpaar ($OPuK$ - „Outer Public Key“ und OPK - „Outer Private Key“) gebildet, das sich im Besitz des Users auf dem Security-Token befindet (S* in Abbildung 1). Die Schlüssel auf dem Security-Token sind nur nach Eingabe des korrekten PINs zugänglich, womit eine „Zwei-Faktor-Authentifizierung“ gewährleistet ist. Der innere symmetrische Schlüssel (ISK, verschlüsselt mit dem inneren öffentlichen Schlüssel), der innere private Schlüssel (IPK,

verschlüsselt mit dem äußeren öffentlichen Schlüssel), und der innere öffentliche Schlüssel ($IPuK$) bilden die innere Hülle, die Rechteschicht (K* in Abbildung 1). Mit dem inneren symmetrischen Schlüssel sind wiederum die Pseudonyme verschlüsselt [1] , welche mit den eigentlichen medizinischen Patientendaten (CD in Abbildung 1) die innerste Hülle darstellen, die pseudonymisierte Datenschicht. Mehrere Pseudonyme, die mit demselben Datensatz referenziert sind, entsprechen einzelnen Datenzugriffsautorisierungen. Abbildung 1 stellt auch die HSM-Interaktionen bezüglich der kryptografischen Operationen dar: Während die kryptografischen Funktionen des Security-Tokens nur für die Client-seitigen Authentifizierungsschritte und Ver-/Entschlüsselung mit dem Session-Schlüssel (OSK_S - „Outer Symmetric Key“ für die Verschlüsselung von ansonsten unverschlüsselten Daten für den Transport zwischen Client und Server) benötigt werden, werden in dem HSM die kryptografischen Hauptoperationen durchgeführt (vgl. Abschnitt 4). Im Massenpseudonymisierungsszenario unterstützt das Hüllenmodell verschiedene Typen von Usern (Rollen) einschließlich den des Dateneigentümers (Datenadministrator oder Patient), eines (optionalen) Autorisierten, des Operators, des (PIPE) Administrators (nicht abgebildet) und des sekundären Users:

Dateneigentümer Der Dateneigentümer hat die volle Kontrolle über seine Unterlagen und kann nach eigenem Ermessen Daten hinzufügen oder entfernen sowie Datenzugriffsautorisierungen definieren und diese auch, falls nötig, widerrufen. Im klassischen PIPE-Szenario agiert der Patient selbst als Dateneigentümer, was im Zuge einer Massenpseudonymisierung jedoch nicht optimal wäre, da hierfür die aktive Teilnahme einer großen Anzahl an Patienten am Pseudonymsierungsprozess aufgrund der großen Menge an medizinischen Daten notwendig wäre. Daher agiert eine Gruppe von Datenadministratoren während der Pseudonymisierungsprozedur als (Interims-) Eigentümer, die denselben inneren symmetrischen Schlüssel einer „Haupteigentümerinstanz“ (K_{MO} des „Master Data Owner“ in Abbildung 1) verwendet, welcher für jeden Datenadministrator mit seinem jeweiligen inneren öffentlichen Schlüssel (K_{DA} in Abbildung 1) chiffriert ist.

Autorisierter Optional können medizinische Daten für vertrauenswürdige Dritte freigegeben werden, wobei diese dann die Rolle des Autorisierten verkörpern. Für jeden Autorisierten wird ein neues Pseudonym generiert, das auf das medizinische Dokument verweist und auch dem Dateneigentümer bekannt ist. Durch das Löschen dieses Pseudonyms kann der Dateneigentümer die Autorisierung widerrufen.

Administrator Im Gegensatz zu konventionellen Datenstrukturen hat der (System-) Administrator hier keinerlei Zugangsrechte zu den Patientenunterlagen (oder genauer: er kann nicht die Gesamtheit aller medizinischen Unterlagen eines bestimmten Patienten einsehen, da ihm die Informationen für die Wiederherstellung der Verbindung zwischen dem Patient und dessen Unterlagen unbekannt sind), sondern er kümmert sich nur um administrative Aufgaben wie die Verwaltung der User-Instanzen einschließlich der Vergabe der Security-Token an die User.

Operator Die einzige Aufgabe der Operatoren ist es, bestimmte Shares der inneren privaten Schlüssel der User aus Backupgründen entgegenzunehmen und „aufzubewahren“, indem

[1] Im Allgemeinen könnten die Pseudonyme direkt mit dem inneren öffentlichen Schlüssel chiffriert werden und wären gegen unautorisierten Zugriff geschützt. Dennoch ist die Verwendung eines zusätzlichen inneren symmetrischen Schlüssels von Vorteil: Einerseits können symmetrische Verschlüsselungen schneller ausgeführt werden als mit zeitaufwendigeren asymmetrischen Verfahren. Andererseits hält dies den User davon ab, direkten Zugang zum inneren symmetrischen Schlüssel zu erhalten, da dieser nur innerhalb der sicheren Umgebung eines serverseitigen HSMs, wo die Pseudonyme ver- und entschlüsselt werden, im Klartext vorliegt.

die Schlüsselteile mit deren individuellen inneren symmetrischen Operatorschlüsseln chiffriert und gespeichert werden. Bei Verlust eines Security-Tokens kann nicht mehr auf den wichtigen inneren privaten Schlüssel des Users zugegriffen werden, wobei eine vorbestimmte Anzahl an Operatoren zusammenarbeitet, um diesen Schlüssel zu rekonstruieren, der ansonsten nicht mehr verfügbar wäre (aufgrund des Verlustes des äußeren privaten Schlüssels).

Sekundärer User Der sekundäre User (z.B. eine Forschungseinrichtung) hat die Genehmigung, medizinische Daten für Forschungszwecke einzusehen, kann allerdings nicht die Verbindung zwischen medizinischen Daten und den persönlichen Daten des entsprechenden Patienten herstellen. Da die Daten bereits pseudonymisiert gespeichert sind, wird dem sekundären User direkter Zugang zu den (medizinischen) Daten gewährt. Der Zugang wird mitgeloggt, um den Patienten über den Datenzugriff zu informieren.

Die Patientenunterlagen werden nicht direkt über die einzigartigen Datensatz-Identifikatoren abgerufen, sondern mittels der Pseudonyme. Der Datenzugriff wird über die Kenntnis (kryptografischer Schlüssel) der korrekten referenzierten Pseudonyme reguliert. Für den Fall, dass ein optionaler Autorisierter (Rolle) definiert wird, müssen die verschiedenen Zugriffsberechtigungsstufen von Dateneigentümer und Autorisierten über verschiedene Typen von Pseudonymen umgesetzt werden: Root- und Shared-Pseudonyme.

Root-Pseudonyme Root-Pseudonyme ($oPSN$) werden nur für den Dateneigentümer generiert, der als einzige Person diese Pseudonyme kennt. Sie sind die primären Referenzen zu den Daten des Eigentümers. Jeder Datensatz ist mit genau einem Root-Pseudonym verbunden, da pro Patientendatensatz nur ein Dateneigentümer (Haupteigentümerinstanz) erlaubt ist.

Shared-Pseudonyme Shared-Pseudonyme ($aPSN$) werden für jede einzelne Autorisierung generiert, d.h. für jede Eigentümer/Autorisierter/Patientendatensatz-Kombination. Wie der Name andeutet, sind diese Pseudonyme sowohl dem Dateneigentümer als auch dem Autorisierten bekannt, wobei der Autorisierte auf die Shared-Pseudonyme angewiesen ist, um die Daten des Eigentümers abzurufen, während der Eigentümer den Zugang zu seinen Daten kontrolliert, indem er diese Pseudonyme erstellt und löscht.

Um die pseudonymisierten identifizierenden und medizinischen Datensätze logisch zu trennen, wird ein weiterer Satz an Pseudonymen eingeführt: Identification- und Health-Pseudonyme.

Identification-Pseudonym Identification-Pseudonyme (PSN_{ID}) beziehen sich auf persönliche, den Patienten identifizierende Datensätze, die beispielsweise den vollständigen Namen, Adresse und weitere nicht-medizinische Informationen enthalten.

Health-Pseudonyme Health-Pseudonyme (PSN_{HE}) sind mit den eigentlichen Gesundheitsdaten des Patienten referenziert.

Die Identification- und Health-Pseudonyme gehen eine 1:1-Parent/Child-Beziehung ein, um die Verbindung zwischen Patient und medizinischen Daten darzustellen. Damit diese Verbindung für nicht autorisierte Personen nicht zu erkennen ist, werden die Pseudonyme, wie schon erwähnt, mit dem inneren symmetrischen Schlüssel chiffriert:

$$\left\{\{PSN_{ID}, PSN_{HE}\}_{ISK_{MO}}\right\} \tag{1}$$

Da die Pseudonyme selbst keine semantische Information beinhalten, unterstützt das PIPE-System sogenannte Keywords, die mit den Pseudonymen referenziert werden, um einen Ab-

fragemechanismus für bestimmte medizinische Daten zu realisieren. Während unstrukturierte (beliebige) Keywords für Bereichsabfragen schlecht geeignet sind und außerdem womöglich zu viel Information enthalten könnten, erlaubt PIPE nur strukturierte Keywords, die nach speziellen Templates konstruiert werden. Diese Vorlagen entsprechen verschiedenen Eigenschaften von Patientendaten wie Dokumententyp (z.B. Anamnese, Computertomografie), Krankheitstyp und Datum. Internationale Standards wie die International Statistical Classification of Diseases and Related Health Problems (ICD) oder die Logical Observation Identifiers Names and Codes (LOINC) sind besonders als Templates für Keywords geeignet. Um die Relation zwischen Pseudonym und Keyword zu verbergen, wird jedem (konstruierten) Keyword ein einzigartiger Identifikator zugewiesen, der wiederum verschlüsselt mit dem Pseudonym referenziert ist. Die Verschlüsselung dieses Identifikators statt dem Schlüsselwort selbst gestattet es anderen Benutzern, dasselbe Keyword zu verwenden (z.B. für Root- und Shared-Pseudonyme, die demselben Gesundheitsdatensatz zugewiesen sind). Im folgenden Abschnitt werden die wichtigsten Operationen der Massenpseudonymisierung erklärt, einschließlich der Authentifizierung der Datenadministratoren, der Prozedur der Massenpseudonymisierung selbst und der darauffolgende Transfer der Eigentümerrechte an den medizinische Daten zu den entsprechenden Patienten. Damit der Prozess der Massenpseudonymisierung weitgehend automatisiert abläuft, müssen im Hinblick auf die Struktur der medizinischen Daten folgende Bedingungen erfüllt sein:

- Die medizinischen Daten liegen bereits in digitaler Form vor.
- Die Daten sind entweder schon in identifizierende und medizinische Daten getrennt oder so strukturiert, dass die Aufteilung automatisch durchgeführt werden kann, z.B. codiert im HL7 CDA[2] Format.
- In den medizinischen Datensätzen sind beschreibende Information so strukturiert, dass sie automatisch als Keywords extrahiert werden können und auch den Templates entsprechen.

4 Operationen der Massenpseudonymisierung

Dieser Abschnitt beschreibt die primären Szenarien einschließlich der Nachrichten, die zwischen dem Client- (User) und den Server-Komponenten von PIPE ausgetauscht werden. Im Folgenden wird die verwendete Notation vorgestellt.

4.1 Notation

Es bezeichnen U, S, und D die User-, Server- und Datenbank-Instanzen und $sender \rightarrow receiver : message$ die ausgetauschten Nachrichten. Es bezeichnet $S \xrightarrow{Ret.} D : query \Rightarrow response$ eine Datenabfrageprozedur („Retrieval“) mit einer Abfrage und der entsprechenden (erwarteten) Antwort und $S \xrightarrow{St.} D : item\ (\Rightarrow response)$ eine Datenspeicherungsprozedur („Storage“) mit dem zu speichernden Datensatz und einer möglichen Antwort. Es bezeichnet $S : f_1(x_1)(\rightarrow f_2(x_2), ...)$ eine Prozedur, die auf dem Server ausgeführt wird, mit möglichen Folgeoperationen. Es bezeichnen $E_{T \vee H}(key, item) = \{item\}_{key}$ und $D_{T \vee H}(key_1, \{item\}_{key_2}) = item$ die Ver- und Entschlüsselungsvorgänge („Encryption/Decryption“), die auf dem Security-Token T oder im HSM H ausgeführt werden, mit entsprechenden Schlüsseln und Elementen, wobei $key_1 = key_2$ für symmetrische Schlüssel gilt. $G_H(item)$ bezeichnet die Generierung von Elementen im HSM, einschließlich der Pseudonyme und Challenges. $OPuK_U, OPK_U, IPuK_U, IPK_U \in$

[2] Health Level 7 Clinical Document Architecture

$\mathcal{K}_A$ entsprechen dem äußeren öffentlichen, äußeren privaten, inneren öffentlichen und inneren privaten Schlüssel, wobei die Schlüssel Elemente der Menge der aktuell verwendeten symmetrischen Schlüsseln $\mathcal{K}_A$ sind und durch die Bitstrings $\mathcal{K}_A = \{0,1\}^{l_A}$ (typischerweise $l_A = 2048$ für RSA-Schlüssel) mit den Eigenschaften $OPuK = OPK^{-1}, IPuK = IPK^{-1}$ repräsentiert werden, sodass folgendes gilt:

$$item = D(OPK, E(OPuK, item)) = D(OPuK, E(OPK, item)) \tag{2}$$

$$item = D(IPK, E(IPuK, item)) = D(IPuK, E(IPK, item)) \tag{3}$$

$ISK_U, ISK_S, OSK_S \in \mathcal{K}_S$ entsprechen dem inneren symmetrischen Schlüssel des Users, dem inneren symmetrischen Schlüssel des Servers (Logik/HSM-Schlüssel) und dem äußeren symmetrischen Schlüssel des Servers (Session-Schlüssel), wobei die Schlüssel Elemente der Menge der aktuell verwendeten symmetrischen Schlüsseln $\mathcal{K}_S$ sind und durch die Bitstrings $K_S = \{0,1\}^{l_S}$ (typischerweise $l_S = 256$ für AES-Schlüssel) repräsentiert werden, sodass folgendes gilt:

$$item = D(ISK, E(ISK, item)) \tag{4}$$

$$item = D(OSK, E(OSK, item)) \tag{5}$$

Es bezeichnet $IUID_U \in \mathcal{I}$ den internen Identifikatoren des Users als Element der Menge der aktuell verwendeten Identifikatoren $\mathcal{I}$, die durch die Bitstrings $\mathcal{I} = \{0,1\}^{l_I}$ repräsentiert werden, wobei $l_I = 64$. $oPSN_{ID}, oPSN_{HE} \in \mathcal{P}$ repräsentieren die Root-Identification- und Root-Health-Pseudonyme als Elemente der Menge der aktuell verwendeten Pseudonyme $\mathcal{P}$, die durch die Bitstrings $\mathcal{P} = \{0,1\}^{l_P}$ repräsentiert werden, wobei $l_P = 64$. $\{item\}^+$ bezeichnet die Menge von einem oder mehreren Elementen als Ergebnis einer Abfrage, einschließlich Pseudonymen und Identifikatoren, und $item^C$ bezeichnet ein vorläufig generiertes Element, welches noch auf Einzigartigkeit überprüft werden muss („candidate“). Ein Keyword KWD wird durch die Verkettung von Elementen aus verschiedenen Template-Familien $\mathcal{KT}_1, ..., \mathcal{KT}_n$ generiert, wobei KID dem Keyword-Identifikatoren entspricht und KW_E die Menge der aktuell verwendeten Keywords bezeichnet:

$$KWD = \{k_1||...||k_n \mid k_1 \in \mathcal{KT}_1, ..., k_n \in \mathcal{KT}_n\} \tag{6}$$

Es bezeichnet $data_{IN}$ den zu pseudonymisierenden Input-Datensatz mit dem dazugehörigen Identifikator RID_{IN}, und $data_{ID}$ und $data_{HE}$ bezeichnen die getrennten identifizierenden sowie die medizinischen Datensätze mit ihren Identifikatoren RID_{ID} und RID_{HE}. Und schließlich bezeichnet $U = \{MO, DA, OW\}$ die Rollen wie folgt: Hauptdateneigentümer („Master Owner“), Datenadministrator als (Interims-) Dateneigentümer und Patient als endgültiger Dateneigentümer („Owner“).

4.2 Authentifizierung

Zu Beginn jeder Sitzung muss sich jeder User am System authentifizieren, um einerseits seine Identität nachzuweisen und andererseits dem HSM die für die weiteren Arbeitsschritte nötigen Schlüssel zur Verfügung zu stellen. Das folgende Protokoll ist aus der Sicht eines Datenadministrators beschrieben.

- Der Datenadministrator verschlüsselt seinen Identifikator $IUID_{DA}$ mit dem äußeren öffentlichen Schlüssel des Servers $OPuK_S$ und transferiert ihn zum Server, wo der Identifikator entschlüsselt wird und der entsprechende äußere öffentliche Schlüssel $OPuK_{DA}$

in der Datenbank gesucht wird. Falls der User existiert, wird der Schlüssel retourniert, andernfalls wird das Authentifizierungsprotokoll abgebrochen.

$$U_{DA} \rightarrow S : E_T(OPuK_S, IUID_{DA}) \tag{7}$$

$$S \xrightarrow{Ret.} D : D_H\left(OPK_S, \{IUID_{DA}\}_{OPuK_S}\right) \Rightarrow \tag{8}$$

$$\Rightarrow \begin{cases} OPuK_{DA} & \text{if } IUID_{DA} \in \mathcal{I} \\ \emptyset & \text{if } IUID_{DA} \notin \mathcal{I} \end{cases}$$

- Der Server generiert eine Zufallszahl RV_S, welche mit dem äußerem öffentlichen Schlüssel des Datenadministrators verschlüsselt wird und an diesen gesendet wird.

$$S \rightarrow U_{DA} : E_H\left(OPuK_{DA}, G_H\left(RV_S\right)\right) \tag{9}$$
$$\text{where } RV_S = \{0,1\}^{l_R},\ l_R = 64$$

- Der Datenadministrator dechiffriert diese Challenge mit seinem äußeren privaten Schlüssel OPK_{DA} und retourniert ihn (verschlüsselt mit dem äußeren öffentlichen Schlüssel des Servers).

$$U_{DA} \rightarrow S : E_T(OPuK_S, RV_{DA}) \tag{10}$$
$$\text{where } RV_{DA} = D_T\left(OPK_{DA}, \{RV_S\}_{OPuK_{DA}}\right)$$

- Der Server kontrolliert die vom Datenadministrator erhaltene Zufallszahl RV_{DA} auf ihre Gültigkeit. Falls gültig, wird vom Server dessen verschlüsselter geheimer innerer privater Schlüssel abgefragt und seine Version des inneren symmetrischen Schlüssels des Hauptdateneigentümers ISK_{MO} abgerufen, andernfalls wird das Authentifizierungsprotokoll abgebrochen.

$$S \xrightarrow{Ret.} D : IUID_{DA} \Rightarrow \{IPK_{DA}\}_{OPuK_{DA}}, \{ISK_{MO}\}_{IPuK_{DA}}, IUID_{MO} \tag{11}$$
$$\text{iff } RV_{DA} = RV_S$$

- Der Server generiert dann einen neuen äußeren symmetrischen Schlüssel als Session-Schlüssel und sendet diesen gemeinsam mit dem zu dechiffrierenden inneren privaten Schlüssel an den Datenadministrator.

$$S \rightarrow U_{DA} : \{IPK_{DA}\}_{OPuK_{DA}}, E_H\left(OPuK_{DA}, G_H\left(OSK_S\right)\right) \tag{12}$$
$$\text{where } OSK_S \in \mathcal{K}_S$$

- Der Datenadministrator entschlüsselt sowohl den inneren privaten als auch den Session-Schlüssel und retourniert den inneren privaten Schlüssel, chiffriert mit dem Session-Schlüssel, an den Server, wo er verwendet wird, um den inneren symmetrischen Schlüssel des Hauptdateneigentümers zu dechiffrieren. Für andere Benutzer (z.B. Autorisierte) wird natürlich der eigene innere symmetrische Schlüssel dechiffriert.

$$U_{DA} \rightarrow S : E_T(OSK_S, IPK_{DA}) \tag{13}$$
$$\text{where } OSK_S = D_T\left(OPK_{DA}, \{OSK_S\}_{OPuK_{DA}}\right)$$
$$\text{and } IPK_{DA} = D_T\left(OPK_{DA}, \{IPK_{DA}\}_{OPuK_{DA}}\right)$$

$$S : D_H\left(OSK_S, \{IPK_{DA}\}_{OSK_S}\right) \tag{14}$$

$$S : D_H\left(IPK_{DA}, \{ISK_{MO}\}_{IPuK_{DA}}\right) \tag{15}$$

- Der innere symmetrische Schlüssel des Hauptdateneigentümers und der innere private Schlüssel des authentifizierten Datenadministrators sowie beide Identifikatoren stehen jetzt für weitere Arbeitsschritte zur Verfügung.

$$S : ISK_{MO}, IUID_{MO}, IPK_{DA}, IUID_{DA} \tag{16}$$

4.3 Pseudonymisierung der Daten

Das Pseudonymisieren der Daten ist die zentrale Operation der Massenpseudonymisierung, wobei die medizinischen von den identifizierenden Daten getrennt werden.

- Zuerst ruft der Server die zu pseudonymisierenden Daten ab und trennt sie in separate identifizierende und medizinische Datensätze sowie Keywords.

$$S \xrightarrow{Ret.} D : RID_{IN} \Rightarrow data_{IN} \tag{17}$$

$$S : f_{separate}\left(data_{IN}\right) = \{data_{ID}, data_{HE}, KWD\} \tag{18}$$

- Der Server generiert neue Root-Identification- und Root-Health-Pseudonyme und kontrolliert diese auf deren Einzigartigkeit. Darüber hinaus wird kontrolliert, ob die Keywords schon in Verwendung sind. Falls ja, wird der Keyword-Identifikator abgerufen, andernfalls wird das neue Keyword in der Datenbank gespeichert.

$$S : G_H(oPSN_{ID}^C) \rightarrow f_{check}(oPSN_{ID}^C) \rightarrow \tag{19}$$

$$\rightarrow \begin{cases} oPSN_{ID} = oPSN_{ID}^C & \text{if } oPSN_{ID}^C \notin \mathcal{P} \\ G_H(oPSN_{ID}^C) & \text{if } oPSN_{ID}^C \in \mathcal{P} \end{cases}$$

$$S : G_H(oPSN_{HE}^C) \rightarrow f_{check}(oPSN_{HE}^C) \rightarrow \tag{20}$$

$$\rightarrow \begin{cases} oPSN_{HE} = oPSN_{HE}^C & \text{if } oPSN_{HE}^C \notin \mathcal{P} \\ G_H(oPSN_{HE}^C) & \text{if } oPSN_{HE}^C \in \mathcal{P} \end{cases}$$

$$S : f_{check}(KWD) \rightarrow \begin{cases} KID & \text{if } KWD \in \mathcal{KW}_E \\ \emptyset & \text{if } KWD \notin \mathcal{KW}_E \end{cases} \rightarrow \tag{21}$$

$$\rightarrow S \xrightarrow{St.} D : KWD \Rightarrow KID \quad \text{iff } f_{check}(KWD) = \emptyset$$

- Der Server verschlüsselt die Pseudonyme, den Keyword- und den Hauptdateneigentümer-Identifikatoren mit dessen innerem symmetrischen Schlüssel und speichert diese in der Datenbank. Die pseudonymisierten Datensätze werden getrennt gespeichert und die entsprechenden Identifikatoren mit den Klartext-Pseudonymen referenziert. Falls zu einem Patienten mehrere medizinische Unterlagen zu pseudonymisieren sind, werden die Identifikationsdaten nur einmal gespeichert und alle Identification-Pseudonyme werden mit

dem entsprechenden Datensatz-Identifikator verbunden.

$$S \xrightarrow{St.} D : E_H\left(ISK_{MO}, \{IUID_{MO}, oPSN_{ID}, oPSN_{HE}, KID\}\right) \tag{22}$$

$$S \xrightarrow{Ret.} D : data_{ID} \Rightarrow \begin{cases} RID_{ID} & \text{if } data_{ID} \in D \\ \emptyset & \text{if } data_{ID} \notin D \end{cases} \rightarrow \tag{23}$$

$$\rightarrow S \xrightarrow{St.} D : data_{ID} \Rightarrow RID_{ID} \quad \text{iff } data_{ID} \notin D$$

$$S \xrightarrow{St.} D : data_{HE} \Rightarrow RID_{HE} \tag{24}$$

$$S \xrightarrow{St.} D : \{oPSN_{ID}, RID_{ID}\}, \{oPSN_{HE}, RID_{HE}\} \tag{25}$$

4.4 Transfer der Eigentümerrechte

Nach der Pseudonymisierung durch die Datenadministratoren können die Eigentümerrechte schließlich schrittweise an die Patienten übergeben werden, wobei neue Root-Pseudonyme erstellt werden.

- Mittels identifizierender Information $info_{ID}$ (z.B. Name des Patienten) kann der entsprechende identifizierende Datensatz abgerufen und somit die Root-Identification-Pseudonyme ermittelt werden. Mit dem inneren symmetrischen Schlüssel des Hauptdateneigentümers können in weiterer Folge die Root-Health-Pseudonyme und somit alle Identifikatoren der medizinischen Datensätze abgerufen werden. Weiters werden alle Keyword-Identifikatoren ermittelt.

$$U_{DA} \rightarrow S : E_T\left(OSK_S, info_{ID}\right) \tag{26}$$

$$S \xrightarrow{Ret.} D : D_H\left(OSK_S, \{info_{ID}\}_{OSK_S}\right) \Rightarrow RID_{ID} \tag{27}$$

$$S \xrightarrow{Ret.} D : RID_{ID} \Rightarrow \{oPSN_{ID}\}^+ \tag{28}$$

$$S \xrightarrow{Ret.} D : E_H\left(ISK_{MO}, \{oPSN_{ID}\}^+\right) \Rightarrow \{oPSN_{HE}, KID\}^+_{ISK_{MO}} \tag{29}$$

$$S : D_H\left(ISK_{MO}, \{oPSN_{HE}, KID\}^+_{ISK_{MO}}\right) \tag{30}$$

$$S \xrightarrow{Ret.} D : \{oPSN_{HE}\}^+ \Rightarrow \{RID_{HE}\}^+ \tag{31}$$

- Für jeden medizinischen Datensatz werden neue Identification- und Health-Pseudonym-Kandidaten generiert und auf deren Einzigartigkeit geprüft (vgl. Abschnitt 4.3), die Pseudonyme den Datensatz-Identifikatoren zugewiesen (Root-Identification-Pseudonyme werden offensichtlich demselben identifizierenden Datensatz zugewiesen) und die Pseudonyme und die Keyword-Identifikatoren mit dem inneren symmetrischen Schlüssel des Patienten ISK_{OW} verschlüsselt.

$$S \xrightarrow{St.} D : E_H\left(ISK_{OW}, \{IUID_{OW}, oPSN_{ID}, oPSN_{HE}, KID\}^+\right) \tag{32}$$

$$S \xrightarrow{St.} D : \{oPSN_{ID}, RID_{ID}\}^+, \{oPSN_{HE}, RID_{HE}\}^+ \tag{33}$$

- Schließlich müssen noch die alten Pseudonyme des Hauptdateneigentümers (verschlüsselte und Klartext-Versionen) aus der Datenbank gelöscht werden, um die Eigentumsrechte des Datenadministrators komplett zu widerrufen.

5 Ausblick

Die elektronische Patientenakte (ELGA) ermöglicht die strukturierte und erweiterbare Sammlung medizinischer Daten, die für klinische Forschungsstudien benötigt werden. Somit wird nicht nur die Optimierung klinischer Forschungen ermöglicht, sondern die Resultate weisen aufgrund einer größeren Anzahl an Proben eine höhere statistische Aussagekraft auf. Ein Problem stellt dabei die sich ergebenden Datenschutzbedenken bei elektronisch gespeicherten medizinischen Daten dar, die für eine sekundäre Nutzung zugänglich gemacht werden. Dieser Beitrag präsentiert eine neue Methode zur effizienten Massenpseudonymisierung von Gesundheitsdaten und gibt einen detaillierten Überblick über die Systemarchitektur und die wichtigsten Arbeitsschritte. Dieser Ansatz ermöglicht Forschungsorganisationen den Zugriff auf schon vorhandene medizinische Daten bei gleichzeitiger Wahrung des Datenschutzes.

6 Danksagungen

Diese Arbeit wurde im Rahmen des Kompetenzzentrums Secure Business Austria durchgeführt, das vom Bundesministerium für Wirtschaft und Arbeit (BMWA) sowie der Stadt Wien gefördert wird. Diese Arbeit wurde durch die FIT-IT Forschungsinitiative Trust in IT Systems (Fördervertrag 816158) gefördert.

Literatur

[BC96] BARROWS, Randolph C.; CLAYTON, Paul D.: Privacy, Confidentiality, and Electronic Medical Records. In: *Journal of the American Medical Informatics Association* 13 (1996), S. 139–148

[Cau06] CAUMANNS, Joerg: Der Patient bleibt Herr seiner Daten. In: *Informatik-Spektrum* (2006), S. 321–331

[EG01] ERNST, Frank R.; GRIZZLE, Amy J.: Drug-Related Morbidity and Mortality: Updating the Cost-of-Illness Model. In: *Journal of the American Pharmacists Association* 41 (2001), Nr. 2, S. 192–199

[Eur95] EUROPEAN UNION: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. In: *Official Journal of the European Communities* L 281 (1995), S. 31–50

[FH01] FISCHER-HÜBNER, Simone: *IT-Security and Privacy: Design and Use of Privacy-Enhancing Security Mechanisms*. Springer, 2001

[Fra05] FRAUNHOFER INSTITUT: *Spezifikation der Lösungsarchitektur zur Umsetzung der Anwendungen der elektronischen Gesundheitskarte*. 2005

[HGG05] HORNUNG, Gerrit; GOETZ, Christoph F.-J. ; GOLDSCHMIDT, Andreas J. W.: Die künftige Telematik-Rahmenarchitektur im Gesundheitswesen. In: *Wirtschaftsinformatik* 47 (2005), S. 171–179

[Hin03] HINDE, Stephen: Privacy legislation: a comparison of the US and European approaches. In: *Computers and Security* 22 (2003), Nr. 5, S. 378–387.

[HK03] HÜHNLEIN, D.; KNOSOWSKI, Y.: Aspekte der Massensignatur. In: *DACH Security 2003*, 2003

[Mar05] MARIS, Koen: The Human Factor. In: *Proceedings of Hack.lu, Luxembourg*, 2005

[Pet03] PETERSON, Robert L.: Patent: Encryption System for allowing immediate universal access to medical records while maintaining complete patient control over privacy. In: *US Patent US 2003/0074564 A1* (2003). –

[PK05] PFITZMANN, Andreas; KOEHNTOPP, Marit: Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management - A Consolidated Proposal for Terminology. In: *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2005

[Pom94] POMMERENING, Klaus: Medical Requirements for Data Protection. In: *Proceedings of IFIP Congress, Vol. 2*, 1994, 533-540

[PR04] *Kapitel* Secondary use of the Electronic Health Record via Pseudonymisation. In: POMMERENING, Klaus ; RENG, Michael: *Medical And Care Compunetics 1*. IOS Press, 2004, S. 441–446

[RNG+07] RIEDL, Bernhard; NEUBAUER, Thomas; GOLUCH, Gernot; BOEHM, Oswald; REINAUER, Gert; KRUMBOECK, Alexander: A secure architecture for the pseudonymization of medical data. In: *Proceedings of the Second International Conference on Availability, Reliability and Security*, 2007, S. 318–324

[SAG+05] SCHABETSBERGER, Thomas; AMMENWERTH, Elske; GÖBEL, Georg; LECHLEITNER, Georg; PENZ, Robert; VOGL, Raimund; WOZAK, Florian: What are Functional Requirements of Future Shared Electronic Health Records? In: *Connecting Medical Informatics and Bio-Informatics* (2005), S. 1070–1075

[SSP+01] SCHMIDT, Volker; STRIEBEL, Werner; PRIHODA, Heinz; BECKER, Michael; LIJZER, Gregor D.: Patent: Verfahren zum Be- oder Verarbeiten von Daten. In: *German Patent, DE 199 25 910 A1* (2001). –

[Tai04] TAIPALE, Kim A.: Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy and the Lessons of King Ludd. In: *International Journal of Communications Law & Policy* 9 (2004)

[TGU+05] THIELSCHER, Christian; GOTTFRIED, Martin; UMBREIT, Simon; BOEGNER, Frank; HAACK, Jochen; SCHROEDERS, Nikolai: Patent: Data processing system for patient data. In: *Int. Patent, WO 03/034294 A2* (2005). –

[Tho04] THORNBURGH, Tim: Social engineering: the "Dark Art". In: *Proceedings of the first annual ACM Conference on Information Security Curriculum Cevelopment*, ACM Press, 2004. – ISBN 1–59593–048–5, S. 133–135

[Uni06] UNITED STATES DEPARTMENT OF HEALTH & HUMAN SERVICE: HIPAA Administrative Simplification: Enforcement; Final Rule. In: *Federal Register / Rules and Regulations* 71 (2006), Nr. 32

[U.S96] U.S. CONGRESS: Health Insurance Portability and Accountability Act of 1996. In: *104th Congress* (1996).

[U.S03] U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES OFFICE FOR CIVIL RIGHTS: *Summary of the HIPAA Privacy Rule*. Version: 2003