Elektrotechnik & Informationstechnik (2010) 127/5: 135–142. DOI 10.1007/s00502-010-0733-2

Pseudonymisierung für die datenschutzkonforme Speicherung medizinischer Daten

Th. Neubauer IEEE, J. Heurix, A Min Tjoa, E. R. Weippl IEEE

E-Health erlaubt eine effiziente Kommunikation zwischen Gesundheitsdiensteanbietern (GDA) und somit die bessere Verfügbarkeit medizinischer Daten, wodurch nicht nur die Kosten im Gesundheitswesen reduziert, sondern auch die Qualität der Patientenbehandlung verbessert werden kann. Der wesentliche Nachteil der resultierenden Vernetzung besteht in der zunehmenden Wahrscheinlichkeit unautorisierter Zugriffe auf streng vertrauliche Patientendaten, die beispielsweise zur Diskriminierung der betroffenen Personen oder zu Identitätsdiebstahl führen können. Dieser Artikel gibt einen Überblick über gängige Gefahren für den Datenschutz im Gesundheitswesen, so zum Beispiel genetische Diskriminierung. Die Autoren präsentieren einen Ansatz, der basierend auf der Pseudonymisierung von Daten die Vertraulichkeit der Patientendaten garantiert, jedoch gleichzeitig autorisierten Personen den Zugriff gestattet. Dieser Ansatz ermöglicht die direkte Verwendung medizinischer Daten durch GDAs sowie die datenschutzkonforme indirekte Nutzung (Sekundärnutzung) der Daten, beispielsweise durch Forschungseinrichtungen.

Schlüsselwörter: E-Health; Datenschutz; Pseudonymisierung; Gendaten

Privacy-preserving storage of medical data through pseudonymization.

E-health allows better communication between health care providers and higher availability of medical data leading to reduced costs and better quality of patients' treatments. However, the downside of interconnected systems is the increased probability of unauthorized access to highly sensitive records which could result in serious discrimination against the patient or identity theft. This article gives an overview of actual privacy threats, such as genetic discrimination, and presents a pseudonymization approach that keeps the patient's privacy and data confidentiality but – at the same time – allows access for authorized persons. This allows primary use of the medical records by health care providers and privacy-preserving secondary use by researchers.

Keywords: e-health; privacy; pseudonymization; genetic data

Eingegangen am 14. Jänner 2010, angenommen am 9. Mai 2010

1. Einleitung und Problemstellung: E-Health und der Bedarf für Datenschutz

Das heutige Gesundheitswesen ist gekennzeichnet von Maßnahmen zur Steigerung der Effizienz sowie von Prozessoptimierung mit dem Ziel der Kostenreduktion – ohne jedoch die Qualität der medizinischen Versorgung zu gefährden. Der Begriff ,,E-Health'' wird verwendet, um den Einsatz von Informations- und Kommunikationstechnologie für medizinische Arbeitsabläufe zu beschreiben. E-Health umfasst auch die Verbesserung der Kommunikationsmöglichkeiten zwischen intra- und extramuralen Versorgern. Integrationsvorhaben wie die elektronische Gesundheitsakte (ELGA) bauen auf eine neue technische Infrastruktur auf, die den Austausch digitaler Informationen erleichtert. Der technische Fortschritt der letzten Jahre ermöglicht die Umsetzung derartiger Anwendungen ohne eine zu starke Zunahme der Kosten (*Chaudry et al., 2006*; *Evans et al., 1992*).

Ein wichtiges Thema bei diesen vernetzten Systemen und der damit verbundenen hohen Verfügbarkeit sensibler medizinischer Daten ist die Tatsache, dass zentral erfasste und gespeicherte medizinische Aufzeichnungen vor allem gegen Missbrauch geschützt werden müssen. Wenn der Zugriff auf persönliche medizinische Daten nicht nur durch autorisierte Personen erfolgt, hat dies erhebliche negative Auswirkungen auf den einzelnen Patienten. Zum Beispiel könnte die Offenlegung von sensiblen Daten, wie Drogenmissbrauch oder eine HIV-Infektion, zu Diskriminierung führen. Solche Informationen können ausreichen, um den Versicherungsschutz bei neuen Verträgen oder eine erfolgreiche Bewerbung zu verhindern. Selbst eine erhöhte Wahrscheinlichkeit, im Laufe des Lebens eine schwere Krankheit zu entwickeln, kann genügen, um bei Kranken- oder Lebensversicherungen schlechtere Konditionen zu bekommen. Ein Beispiel ist die *genetische Diskriminierung*, d. h. die Benachteiligung von Menschen, die aufgrund von Genmutationen ein erhöhtes Risiko für eine erbliche Erkrankung haben (*Council for Responsible Genetics, 2001*; *Coalition of Genetic Fairness, 2004*). Es gibt zahlreiche dokumentierte Fälle, in denen die Ergebnisse der so genannten ,,prädiktiven genetischen Tests'' von Versicherungsgesellschaften verwendet wurden, um Risikopersonen von der Versicherung auszuschließen. Diese Personen waren in der Regel vollkommen gesund und zeigten zum Zeitpunkt der Antragstellung keinerlei Anzeichen oder klinische Symptome einer Erkrankung. Hinter jedem Fall steht eine persönliche Geschichte (*Council for*

Neubauer, Thomas, Dipl.-Ing. Mag. Dr., European Society for Quality in Healthcare, Vienna Office, 1030 Wien, Österreich; **Heurix, Johannes, Mag., Tjoa, A Min, O. Univ.-Prof. Dipl.-Ing. Dr.,** Institut für Softwaretechnik und Interaktive Systeme, Technische Universität Wien, Favoritenstraße 9-11, 1040 Wien, Österreich; **Weippl, Edgar R. Privatdozent Dipl.-Ing. Mag. Dr.,** SBA Research, 1040 Wien, Österreich
(E-Mail: eweippl@sba-research.org)

Responsible Genetics, 2001; *Coalition of Genetic Fairness, 2004*): Der siebenjährige Danny wurde genetisch getestet, wobei die Ergebnisse eine Neigung zu Herzkrankheiten zeigten. Obwohl er völlig gesund war und Medikamente einnahm, die sein Herzinfarktrisiko senkten, wurde ihm eine Versicherung mit der Begründung einer ‚seit Geburt bestehenden Erkrankung' verweigert. Auch ist die genetische Diskriminierung bei Bewerbungen und Dienstverhältnissen ein Problem: Einer der wahrscheinlich bekanntesten Fälle ist jener von Gary Avary, einem Angestellten der Burlington-Northern-Santa-Fe-Railroad. Der Arbeitgeber verlangte nach einem Krankenstand eine ärztliche Untersuchung. Die geforderte ärztliche Untersuchung hätte auch einen Gentest beinhaltet, den Gary Avary verweigerte. Es folgten eine angedrohte Kündigung und der Gang vor Gericht, bei dem – vermutlich aufgrund des starken medialen Drucks (vgl. (*Halbert, Ingulli, 2008*; *Charles, 2001*)) – ein Vergleich erreicht werden konnte.

Um den Missbrauch von Gesundheitsdaten zu verhindern, wurden Gesetze wie der Genetic Information Nondiscrimination Act (GINA) (*Congress of the United States of America, 2008*), der Health Insurance Portability and Accountability Act (HIPAA) (*United States Department of Health & Human Service, 2006*) oder die EU-Richtlinie 95/46/EC (*European Union, 1995*) eingeführt. Diese Gesetze regeln den Zugang zu und den Austausch von personenbezogenen Gesundheitsdaten. Um diese Gesetze praktisch umsetzen zu können, sind technische Lösungen notwendig, die die Weitergabe von medizinischen Daten an unbefugte Personen erschweren bzw. unmöglich machen. Gleichzeitig sollten große Mengen von digitalisierten Daten aus dem Gesundheitswesen für eine sekundäre Nutzung zur Verfügung stehen, d.h., die indirekte (sekundäre) Verwendung personenbezogener Gesundheitsdaten bspw für die Forschung und Qualitätssichorung (*Safran et al., 2007*).

Der Zugang zu diesen Daten kann dazu beitragen, das Wissen über Krankheiten und Behandlungen zu erweitern sowie die Effektivität und Effizienz der Gesundheitsversorgung zu steigem, was wiederum die Betreuung von Patienten verbessert. Bedenkt man jedoch Berichte (*Stern*, *2008*) über den Handel mit nicht anonymisierten Patienten- und Versicherungsdaten, die von der medizinischen Industrie – ohne die ausdrückliche Zustimmung der Patienten oder Ärzte – erfolgten, so muss der Informationsfluss bei der Weiterverwendung dieser Daten streng kontrolliert werden, um die Rechte der Patienten auf ihre Privatsphäre zu wahren. Die primäre und sekundäre Verwendung von medizinischen Daten zu ermöglichen und gleichzeitig die Privatsphäre von Patienten zu schützen, ist eine große Herausforderung.

1.1 Anonymisierung und Verschlüsselung

Zwei Methoden werden oft als Möglichkeit zur Wahrung der Privatsphäre genannt: Anonymisierung und Verschlüsselung. Anonymisierung zielt darauf ab, die identifizierenden Bestandteile medizinischer Daten zu entfernen, so dass von den Datensätzen nicht mehr auf den Patienten geschlossen werden kann (*Thomson et al., 2005*). Ein erster Schritt zu Anonymisierung kann das Entfernen der Verknüpfung zu Namen und anderen identifizierenden Daten sein. Fischer-Hübner (*Fischer-Hübner, 2001*) definiert Anonymität so, dass der Patient nicht oder nur mit erheblichem Aufwand (Zeit, Kosten, Arbeit) identifiziert werden kann. ‚*k*-anonymity' (*Aggarwal, 2005*) bedeutet, dass neben der Person, von der die Daten stammen, noch mindestens $k-1$ andere Personen existieren, auf die bestimmte Charakteristika ebenfalls zutreffen. Das Individuum ist somit nicht vollständig anonym, aber innerhalb einer Gruppe mit *k* Mitgliedern nicht unterscheidbar.

Ein offensichtlicher Nachteil der Anonymisierung ist, dass sie per definitionem nicht umkehrbar ist. Aus diesem Grund können in Fällen, bei denen die Verknüpfung zur Person zu einem späteren Zeitpunkt (z. B. zur Abrechnung oder zur Übermittlung lung von Studienergebnissen) möglich sein muss, die Daten nicht anonymisiert werden. Ebenso ist eine anonyme Speicherung von Gendaten schwer möglich, da diese Daten per se identifizierende Information beinhalten. Der zweite Standardansatz ist, die Daten zu verschlüsseln und so die Geheimhaltung zu gewährleisten. Eine einfache Verschlüsselung garantiert die Vertraulichkeit der Daten, solange der Schlüssel geheim gehalten wird, nicht in falsche Hände gelangt und keine unverschlüsselte Kopie der Daten existiert. Im Gegensatz zur Anonymisierung ist eine Verschlüsselung umkehrbar, d.h. die Daten können wieder entschlüsselt werden. Bedenkt man die heterogene Systemlandschaft von Spitälern, niedergelassenen Ärzten und Sozial- und Krankenversicherungen, wird die Schlüsselverwaltung von kryptographischen Schlüsseln zu einem komplexen Problem.

1.2 Pseudonymisierung als potentielle Lösung

Pseudonymisierung ist ein Verfahren, das es erlaubt, die Verknüpfung von medizinischen Daten mit den identifizierenden Patienteninformationen nur unter bestimmten, festgelegten und kontrollierten Bedingungen offenzulegen. Die ,,Nichtverknüpfbarkeit'' wird dadurch erreicht, dass die identifizierende Information durch ein Pseudonym ersetzt wird und die Zuordnung vom Pseudonym zur identifizierenden Information durch ein Geheimnis (z. B. einen kryptographischen Schlüssel) geschützt ist. Pseudonymisierung kombiniert die Vorteile der Anonymisierung mit jenen der Vollverschlüsselung. Von den Daten selbst kann nicht auf den Patienten geschlossen werden, allerdings kann im Notfall die Entkoppelung von medizinischen Daten und identifizierender Information, durch spezielle Autorisierung wie z. B. ein Vier-Augen-Prinzip, wieder aufgehoben werden. Pseudonymisierungsverfahren ermöglichen sowohl die (identifizierte) Primärnutzung der Daten im Behandlungsfall als auch die (pseudonymisierte) Sekundärnutzung beispielsweise im Rahmen der Forschung.

Tabelle 1 stellt die Schwierigkeiten zwischen der Wahrung der Privatsphäre des Patienten und der leichten Verwendbarkeit der Daten als Kompromiss zwischen Datenschutz und Transparenz dar. Beide Verfahren, Anonymisierung und Verschlüsselung, verlagern den Schwerpunkt zugunsten der Privatsphäre. erschweren jedoch den direkten Zugriff und die Sekundärverwendung. Pseudonymisierung hat im Vergleich zur Verschlüsselung geringere ,,unerwünschte Nebenwirkungen''.

Tabelle 1. Kompromiss zwischen Schutz der Privatsphäre und Transparenz

Verfahren	Privatsphäre	Transparenz	Bemerkung
Anonymisierung	+	−	identifizierende Information wird entfernt; nicht umkehrbar
Verschlüsselung	+	−	Daten vollständig verschlüsselt; umkehrbar
Pseudonymisierung	+	+	identifizierende Information durch Pseudonym ersetzt; umkehrbar unter definierten Rahmenbedingungen

2. Methodik und Diskussion: Das PIPE-Modell

Das PIPE-Modell (Pseudonymization of Information for Privacy in E-Health) ermöglicht die sichere und datenschutzkonforme Speicherung und Verwaltung sensibler medizinischer Daten. PIPE basiert auf dem Prinzip, dass depersonalisierte medizinische Datensätze, die in entsprechender Häufigkeit auftreten, nicht ausreichen, um einen einzelnen Patienten eindeutig zu identifizieren. Diesen medizinischen Datensätzen werden Pseudonyme zugewiesen. Identifizierende Daten wie Name, Adresse, Telefonnummer etc. werden getrennt von den medizinischen Datensätzen gespeichert und ebenfalls mit einem Pseudonym versehen.

Im Gegensatz zu anderen Pseudonymisierungsansätzen wählt PIPE die Pseudonyme zufällig und leitet sie nicht von anderen Entitäten (z. B. dem Gesundheitsdatensatz) ab. PIPE unterscheidet zwischen Identifikations- und Gesundheitspseudonymen. Zur Steigerung der Sicherheit bilden diese Pseudonyme eine 1:1-Relation, statt mehrere Gesundheitspseudonyme einem Identifikationspseudonym zuzuweisen. Die Pseudonyme sind Teil des Zugriffsmechanismus und deshalb verschlüsselt. Die Kenntnis der korrekten Pseudonyme sowie des kryptographischen Schlüssels erlaubt es dem Benutzer, die Verbindung zwischen Identifikations- und Gesundheitspseudonymen herzustellen.

Das PIPE-Modell basiert auf folgenden Grundsätzen:

- datenschutzkonforme Speicherung: Existierende Ansätze pseudonymisieren die Daten oftmals erst beim Export. PIPE hingegen trennt Gesundheitsdatensätze und Identifikationsdaten bereits beim Speichern. Auf diese Weise sind die Datensätze auch gegen etwaige interne Angreifer geschützt, die direkten Zugriff auf die Datenbank haben (z. B. Administrator). Dieses Konzept stellt einen wesentlichen Unterschied zu alternativen Ansätzen dar, bei denen die Administratoren üblicherweise Kenntnis über die Datenbankstruktur, den Inhalt sowie die logischen Verknüpfungen haben.
- datenschutzkonforme sekundäre Nutzung: Die separate Speicherung erlaubt die sekundäre Nutzung, beispielsweise durch Forschungseinrichtungen, ohne dass eine zusätzliche Anonymisierung der Daten erforderlich ist. Dennoch besteht für autorisierte Nutzer die Möglichkeit, die Verbindung der Daten herzustellen und auf diese Weise die Daten der primären Nutzung zuzuführen.
- patientenzentriertes Autorisierungsmodell: PIPE definiert den Patienten als Dateneigentümer. Der Patient hat zu jedem Zeitpunkt die volle Kontrolle über seine Daten und ist die einzige Person, die Autorisierungen für den Zugriff auf seine Daten an vertrauenswürdige Personen vergeben darf.
- sichere Authentifizierung: Die alleinige Verwendung von Passwörtern ist keine ausreichende Maßnahme zur Absicherung sensibler Anwendungen. Aus diesem Grund verwendet PIPE Smartcards (mit einem integrierten kryptographischen Prozessor) für die Authentifizierung. Der private Schlüssel wird auf der Smartcard generiert und verlässt diese zu keinem Zeitpunkt. Die Smartcard dient zusätzlich als vertrauenswürdiges kryptographisches Modul (z. B. für Ver- und Entschlüsselungsvorgänge).
- kryptographische Standards: PIPE verwendet standardisierte kryptographische Protokolle (z. B. RSA und AES), die bei Bedarf auf einfache Weise ersetzt werden können.

2.1 Benutzerfunktionen

Die Hauptakteure des Systems sind der Patient, die Gesundheitsdiensteanbieter sowie Verwandte und ermächtigte Stellvertreter.

2.1.1 Patient

Der Patient ist der einzige Teilnehmer mit vollständigem Zugriff auf seine Daten. Als Dateneigentümer darf er partielle und vollständige Autorisierungen vergeben und wieder entziehen. Partielle Autorisierungen werden üblicherweise für Gesundheitsdiensteanbieter verwendet und beinhalten den Zugriff auf bestimmte Datensätze. Vollständige Autorisierungen sind für die Vergabe an nahe Verwandte gedacht und ermöglichen dem Autorisierten den Zugriff auf alle Datensätze des Patienten. Der Datenzugriff durch den Patienten erfolgt unter Verwendung von so genannten Root-Pseudonymen (vgl. Abb. 3), die ausschließlich dem Dateneigentümer bekannt sind. Für jeden Datensatz existiert genau ein Root-Health-Pseudonym sowie ein Root-Identifikations-Pseudonym.

2.1.2 Gesundheitsdiensteanbieter (GDA)

Der GDA wird vom Patienten zum Zugriff auf bestimmte Datensätze autorisiert. Dieser Vorgang wird durch die Verwendung so genannter ,,Shared-Pseudonymen'' ermöglicht. Analog zu der Verwendung von Root-Pseudonymen werden bei der Autorisierung für den GDA ein Shared-Health-Pseudonym und ein Shared-Identifikations-Pseudonym generiert, die auf den Gesundheitsdatensatz und den Identifikationsdatensatz des Patienten referenzieren. Wie der Name bereits impliziert, sind Shared-Pseudonyme sowohl dem Patienten als auch dem GDA bekannt. Die Pseudonyme sind mit den geheimen Schlüsseln beider Akteure codiert, um einen vertraulichen Zugriff zu garantieren (vgl. Abb. 2). Für jede Autorisierung (auch bei der Autorisierung eines bereits autorisierten GDAs auf einen neuen Datensatz) wird ein neues Paar von Shared-Pseudonymen generiert. Der Patient kann die Autorisierung widerrufen, indem er die Shared-Pseudonyme löscht. Bei Bedarf hat der Patient die Möglichkeit, den GDA (z. B. Hausarzt) für das Hinzufügen neuer Datensätze zu autorisieren.

2.1.3 Verwandter

Im Gegensatz zum GDA kann eine verwandte Person dazu berechtigt werden, auf alle Datensätze des Patienten zuzugreifen. Bei diesem Vorgang erteilt der Patient seinem Verwandten Zugriff auf einen geheimen Schlüssel, der es diesem ermöglicht, die Root-Pseudonyme des Patienten zu entschlüsseln und so Zugriff auf die Gesamtheit seiner Datensätze zu erhalten, indem er praktisch die Rolle des Patienten einnimmt. Mit diesem Schlüssel ist dem Verwandten auch der Zugriff auf Datensätze möglich, die zukünftig durch den Patienten selbst oder den GDA hinzugefügt werden. Die Datenstruktur von PIPE erlaubt in Abhängigkeit der vergebenen Autorisierungen zwei unterschiedliche ,,Sichten'' (vgl. Abb. 1).

Die linke Seite in Abb. 1 zeigt die Sicht auf die Daten, wie sie Administratoren und sekundären Nutzern bzw. internen und externen böswilligen Benutzern zur Verfügung steht. Obwohl die Identifikationsdatensätze und die Gesundheitsdatensätze einsehbar sind, gibt es für diese Benutzer zu keinem Zeitpunkt die Möglichkeit, eine Verbindung zwischen diesen Entitäten herzustellen. Die einzige Möglichkeit würde im Erraten (z. B. Brute-Force) von Zusammenhängen bestehen. Autorisierte Nutzer, d.h. der Patient, der GDA oder Verwandte, können die korrekte Verbindung zwischen Identifikationsdatensatz und den Gesundheitsdatensätzen herstellen. Auf der rechten Seite von Abb. 1 ist die Zusammengehörigkeit zwischen den vier hervorgehobenen Gesundheitsdatensätzen und deren Besitzer, dem Patienten, unter Verwendung des Identifikationsdatensatzes (in der Mitte der Abbildung) ersichtlich.

2.2 Sicherheitsarchitektur

Die Sicherheitsarchitektur von PIPE beruht auf der Verwendung eines Stufenmodells. Jede dieser Stufen ist für die Umsetzung defi-

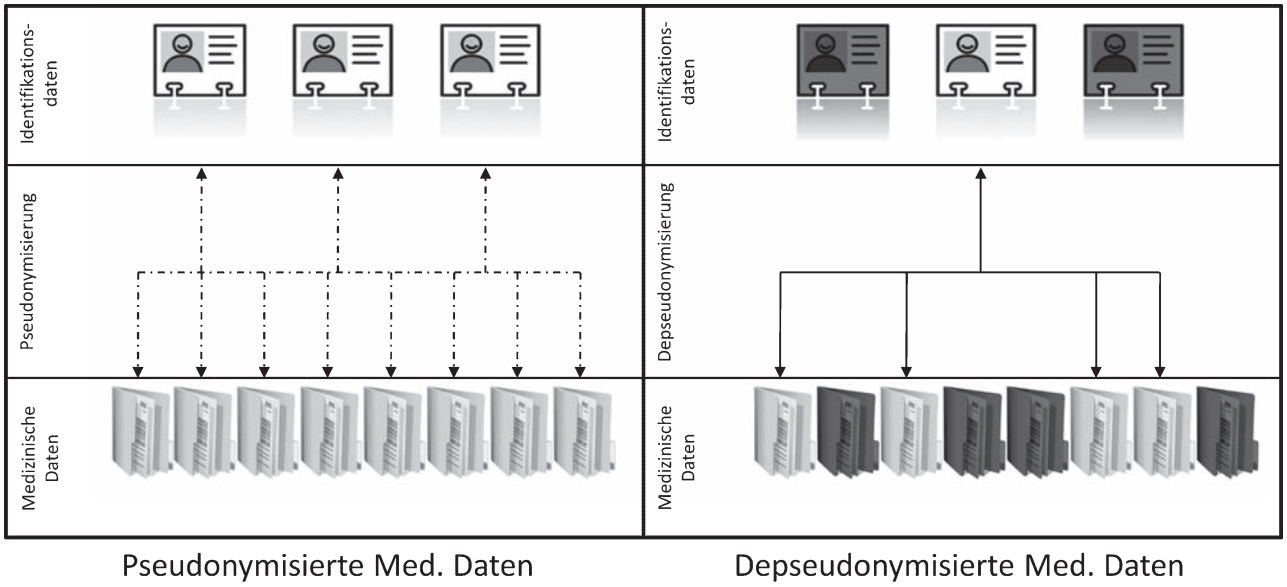

Abb. 1. Ansichten für autorisierte und unautorisierte Personen

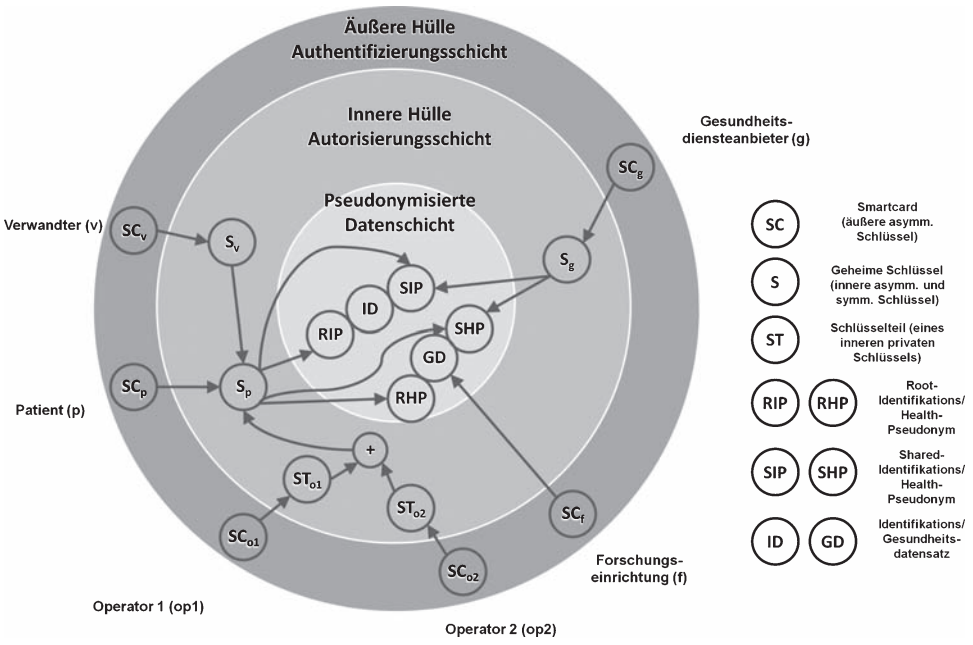

Abb. 2. PIPE-Hüllenmodell

nierter Sicherheitsaspekte verantwortlich (vgl. Abb. 2): Die Authentifizierungsschicht wird durch die Verwendung des äußeren asymmetrischen Schlüsselpaars umgesetzt. Diese Schicht garantiert, dass ausschließlich authentifizierte Benutzer Zugriff auf die zweite Modellschicht erhalten. Der äußere private Schlüssel wird auf der Smartcard generiert und verlässt zu keinem Zeitpunkt die sichere Umgebung der Smartcard. Für die Authentifizierung kann PIPE auch mit etablierten Identitätssystemen wie der Bürgerkarte oder der e-Card integriert werden.

Der Zugriff auf den Schlüssel wird nur nach Eingabe eines PINs gewährt, der ausschließlich dem Besitzer der Smartcard bekannt ist (Zwei-Faktoren-Authentifizierung). Durch Verwendung des äußeren privaten Schlüssels erfolgt der Zugriff auf die nächste Schicht, die Autorisierungsschicht. Dabei wird der äußere private Schlüssel zur Entschlüsselung des inneren privaten Schlüssels herangezogen, der wiederum der Entschlüsselung des inneren symmetrischen Schlüssels dient. Der innere symmetrische Schlüssel ermöglicht dem Benutzer den Zugriff auf die innerste Schicht, die pseudonymisierte Datenschicht. Durch die Entschlüsselung der Pseudonyme mit dem inneren symmetrischen Schlüssel kann der Benutzer die Verbindung zwischen dem Gesundheitsdatensatz und dem dazugehörigen Identifikationsdatensatz herstellen. Sekundärnutzer wie Forschungseinrichtungen erhalten direkten Zugriff auf die depersonalisierten und pseudonymisierten Gesundheitsdatensätze, können jedoch die dazugehörigen Patienten nicht identifizieren. Für den Fall, dass die Smartcard zerstört oder verloren wird, kann eine Gruppe von Operatoren den inneren privaten Schlüssel des Kartenbesitzers mithilfe von zuvor berechneten Teilen des Schlüssels (Secret Sharing) wie-

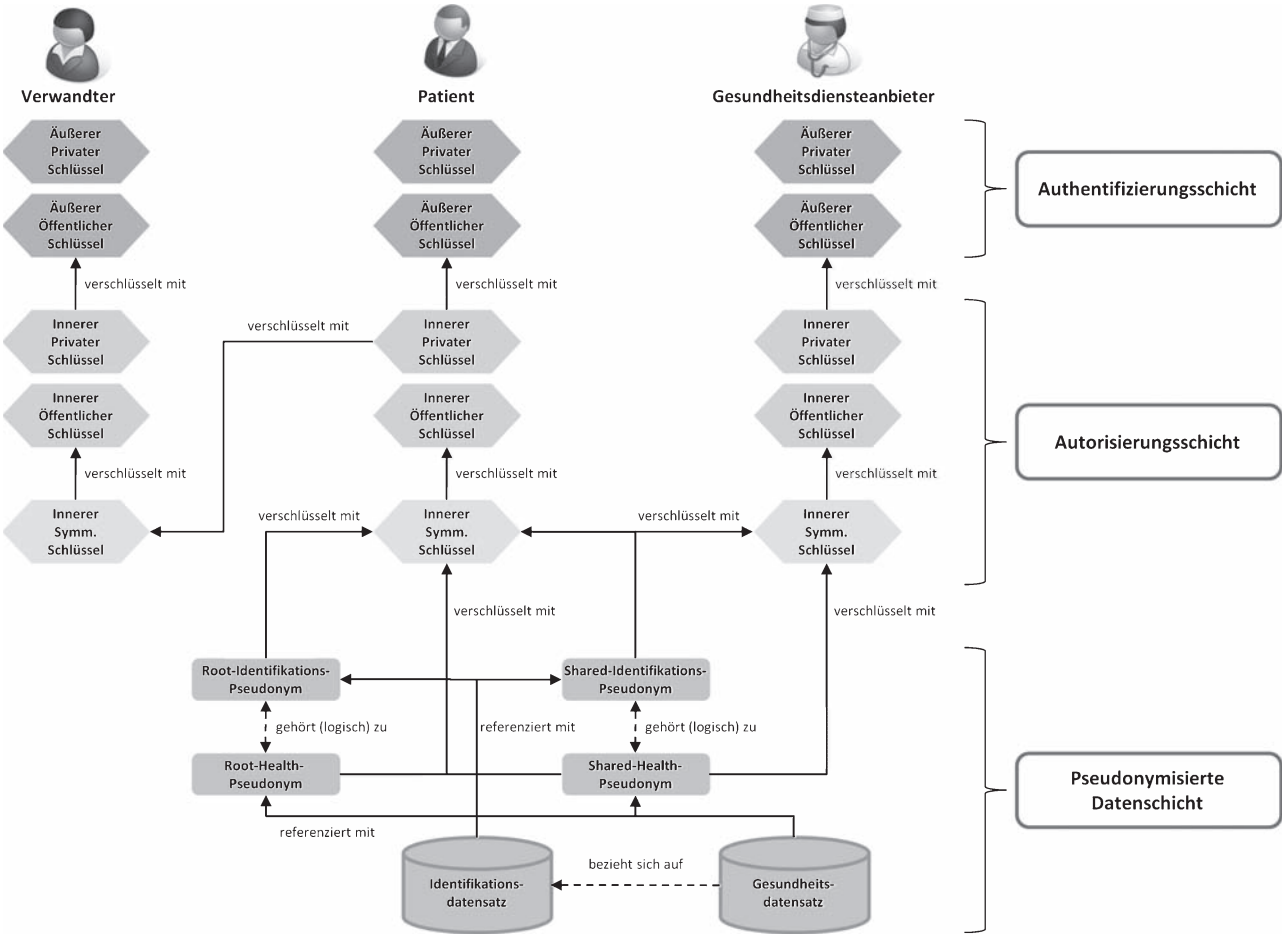

Abb. 3. PIPE-Sicherheitsarchitektur

derherstellen. Der Zugriff auf Daten in einem Notfall kann in PIPE grundsätzlich über ein Vier-Augen-Prinzip erfolgen. In der Praxis ist der viel diskutierte Notfallzugang nur eingeschränkt relevant, da es im Notfall auf jede Sekunde ankommt und die Abfrage von Daten – selbst wenn diese ungesichert vorhanden sind – nicht praktikabel ist.

Abbildung 3 zeigt auf der rechten Seite einen Gesundheitsdiensteanbieter, der für den Zugriff auf einen bestimmten Gesundheitsdatensatz autorisiert wird. Die strichlierte Linie zeigt die Verbindung des Gesundheitsdatensatzes mit dem Identifikationsdatensatz des entsprechenden Patienten (in der Mitte der Abbildung). Aus technischer Sicht wird diese Verbindung durch Health-Pseudonym und Identifikations-Pseudonym repräsentiert und durch die Verschlüsselung mit dem inneren symmetrischen Schlüssel vor unautorisiertem Zugriff geschützt. Die Verbindung kann nur durch Benutzer hergestellt werden, die sowohl die richtige Smartcard als auch den entsprechenden PIN haben. Die Shared-Pseudonyme sind mit den inneren symmetrischen Schlüsseln des Patienten und des GDAs verschlüsselt. Die Root-Pseudonyme sind nur mit dem inneren symmetrischen Schlüssel des Patienten verschlüsselt und somit zu Beginn nur dem Patienten bekannt. Erst nach der Autorisierung eines Verwandten, die mit der Weitergabe des inneren symmetrischen Schlüssels verbunden ist, erhält dieser auch Zugriff auf die Root-Pseudonyme.

Zur Vermeidung von Identitätsdiebstahl setzen wir die Verwendung von Kartenlesegeräten mit Display voraus. Auf diese Weise kann sichergestellt werden, dass der Patient nur jene Personen autorisiert, die tatsächlich Zugriff auf seine Daten erhalten sollen. Die Verifizierung der Integrität des PIPE-Datenbestands erfolgt über Hashwerte, die mit symmetrischen Verifikationsschlüsseln verschlüsselt werden. Damit kann bei der Datenabfrage überprüft werden, ob eine nichtautorisierte Manipulation stattgefunden hat und beispielweise auch eine falsche Verknüpfung von Daten erkannt werden.

2.3 Suche in pseudonymisierten Datensätzen

Die verschlüsselten Pseudonyme dürfen keine semantischen Informationen über die Datensätze, mit denen sie verbunden sind, enthalten. Um die Suche innerhalb der Datensätze dennoch zu ermöglichen, ist ein spezieller Suchmechanismus erforderlich, der auf der Verwendung von Schlüsselwörtern basiert. Der Mechanismus beruht auf der Verwendung vordefinierter, strukturierter Schlüsselwörter, die in Abhängigkeit der Domäne, in der das System eingesetzt ist, ausgetauscht werden können. Beispielsweise kann im Gesundheitswesen der Einsatz von Schlüsselwörtern erfolgen, die eine grobe Klassifikation von Erkrankungen vorgeben (z. B. auf Basis von Standards wie ,,International Statistical Classification of Diseases and Related Health Problems'' (ICD) oder ,,Logical Observation Identifiers Names and Codes'' (LOINC)).

Als Ergänzung dieser Standards können der Dokumententyp (z. B. Röntgen) oder das Datum der Befunderstellung als Schlüsselwörter eingesetzt werden. Zur Gewährleistung des Datenschutzes muss die Verbindung zwischen den Schlüsselwörtern und den Pseudonymen ebenfalls dem Dateninhaber vorbehalten sein. Aus diesem Grund werden die IDs der Schlüsselwörter – wie die Pseudonyme selbst – mit dem inneren symmetrischen Schlüssel des Dateninhabers verschlüsselt und mit den verschlüsselten Pseudonymen referenziert.

3. Pseudonymisierung genetischer Daten

Sobald genetische Daten geschützt werden sollen, stößt die Pseudonymisierung – ebenso wie die Anonymisierung – an ihre Grenzen. Durch die medizinischen Errungenschaften bei der Entschlüsselung der genetischen Daten des Menschen ist die Analyse genetischer Daten zu einem Massenmarkt geworden (vgl. *www.23andme.com*). Dabei werden insbesondere die Veranlagungen für Krankheiten (prädiktive genetische Tests) und die Auswirkungen von Medikamenten (pharmacogenetics) getestet (vgl. (*Roses, 2000*)). Prädiktive genetische Tests umfassen die Analyse der Einzelnukleotid-Polymorphismen (SNP, engl. Single Nucleotide Polymorphism), d. h. die Variationen einzelner Basenpaare in einem DNA-Strang. Bestimmten SNPs bzw. Kombinationen derselben wird nachgesagt, das Risiko im Hinblick auf bestimmte Erkrankungen zu erhöhen. Durch den Aufbau eines SNP-Profils für eine bestimmte Person und den Vergleich mit existierenden Referenzprofilen kann eine erhöhte Prädisposition für das Auftreten einer bestimmten Krankheit identifiziert werden. Genetische Daten weisen per se die Eigenschaft auf, eine bestimmte Person eindeutig identifizieren zu können. Aus diesem Grund ist die einfache Depersonalisierung genetischer Datensätze keine ausreichende Maßnahme, um die Identität der dazugehörigen Person zu schützen.

Wir unterscheiden zwei Vorgangsweisen zur Adressierung dieses Problems:

- Fragmentierung: Die identifizierenden Informationen in einem Datensatz können durch Fragmentierung in mehrere separate Datensätze reduziert werden, die jeder für sich pseudonymisiert werden. Diese Vorgangsweise ist insbesondere bei prädiktiven genetischen Tests zu empfehlen, da meistens nur eine begrenzte Anzahl von SNPs (und korrespondieren Gensequenzen) analysiert werden, die auf einfache Weise fragmentiert werden können.
- Verschlüsselung: In der klinischen Forschung werden oftmals längere Gensequenzen verwendet als bei der Durchführung von prädiktiven genetischen Tests. Die Fragmentierung würde in diesem Fall ein langwieriger Prozess sein, der eine große Anzahl an einzelnen Autorisierungen benötigt, um den gesamten Datensatz lesbar zu machen. In diesem Fall ist die Verschlüsselung jedes Datensatzes mit nur einem Schlüssel die effizientere Vorgangsweise.

Beide Szenarien erfordern die Erweiterung der weiter oben beschrieben Pseudonymisierungsmethode.

3.1 Record Description Kit (RDK)

Das XML-basierte ,,Record Description Kit''[1] ermöglicht die Suche in verschlüsselten XML-Dokumenten. Diese Dokumente können auf nicht vertrauenswürdigen Datenbanken/Servern gespeichert sein. Dieses Konzept nutzt die Struktur von XML-Dokumenten, um auf bestimmte Teile des Dokuments zugreifen zu können, ohne jedoch das gesamte Dokument entschlüsseln zu müssen. Zu diesem Zweck verwendet das RDK folgende Mechanismen:

3.1.1 Schemastruktur

Die Struktur eines XML-Dokuments wird durch das dazugehörige XML-Schema definiert. Jedes Element, Attribut sowie jeder Knoten des XML-Dokuments ist durch einen eindeutigen Bezeichner definiert, der die strukturelle Semantik des Objekts (Pfadinformation) beinhaltet. Dieser Aufbau ermöglicht die Suche nach bestimmten Teilen des XML-Dokuments und verbessert deren Effizienz, da bestimmte Teile der Abfragen ohne Zugriff auf die Datenbank durchgeführt werden können.

3.1.2 Indexstruktur

Um die Geschwindigkeit wiederholt durchgeführter Abfragen zu erhöhen, werden die Indexstrukturen unter Verwendung einer XPath-ähnlichen Syntax aufgebaut. Diese Strukturen unterstützen die Suche nach genauen Übereinstimmungen, die Bereichssuche oder die Suche nach strukturellen Informationen (z. B. alle Knoten auf einem bestimmten Pfad).

3.1.3 Speicherstruktur

Die Speicherung von XML-Dokumenten erfolgt durch die Fragmentierung in Schlüssel/Wert-Paare, wobei die Schlüssel durch gehashte Bezeichner und die Werte durch verschlüsselte XML-Objekte repräsentiert werden. Der Grad der Fragmentierung hängt vom Umfang der im Dokument beinhalteten Information ab.

Anfragen werden in Form (einfacher) XPath-Anfragen akzeptiert, die übersetzt und verarbeitet werden. Danach wird der korrespondierende Dokumentenabschnitt aus der Datenbank abgerufen und entschlüsselt.

3.2 Suche in fragmentierten oder verschlüsselten Datensätzen

Durch die Verwendung des RDK können die oben angeführten Szenarien der Fragmentierung und Verschlüsselung auf folgende Weise umgesetzt werden:

3.2.1 Fragmentierung

Das RDK kann verwendet werden, um die Fragmentierung zu unterstützen, ohne die Gesundheitsdatensätze zu verschlüsseln. In diesem Fall wird der RDK dazu benutzt, um ein XML-codiertes ,,Verzeichnis'' zu erstellen, das alle Datensätze des Benutzers auflistet, auf die er Zugriff hat (entweder als Dateneigentümer oder als Autorisierter). Der Prozess der Fragmentierung muss gewährleisten, dass die Fragmente nicht den Rückschluss auf einen Patienten erlauben, beispielsweise, indem ein Profil der betreffenden Person erstellt wird. Dennoch muss jedes Fragment genug Information beinhalten, um eine sinnvolle Auswertung im Rahmen der klinischen Forschung zu ermöglichen. Jedes Fragment ist mit einem oder mehreren Pseudonymen assoziiert (abhängig von der Anzahl der Autorisierungen), die zusammen mit den entsprechenden Beschreibungen ebenfalls im XML-Dokument gespeichert sind. Diese Vorgangsweise erlaubt die Verwendung von genaueren Beschreibungen anstatt einzelner Schlüsselwörter und garantiert zur gleichen Zeit die Einhaltung des Datenschutzes und der Vertraulichkeit.

3.2.2 Verschlüsselung

Bei der vollständigen Verschlüsselung der Datensätze kann das RDK zur direkten Suche nach bestimmten Teilen innerhalb der verschlüsselten Datensätze herangezogen werden. Zu diesem Zweck müssen die Datensätze mit einer XML-konformen Datenstruktur codiert werden (z. B. unter Verwendung des Health Level 7 Clinical Document Architecture Standard (HL7 CDA)). Diese Spezifikation separiert das Dokument in einen Kopfteil, der für den effizienten Datenaustausch standardisiert ist, und einen Hauptteil, der frei definiert werden kann. Speziell für die Verwendung mit genetischen Daten wurde HL7 Clinical Genomics (HL7 CG) entwickelt, dass die Charakteristika genetischer Daten wie DNA-Sequenzen, SNPs oder individuelle Allele abbilden kann.

[1] Das Record Description Kit wurde in enger Zusammenarbeit mit der ‚Data und Knowledge Engineering'-Gruppe an der Johannes Kepler Universität Linz entwickelt und basiert auf der semantischen XML-Dokumenten-Architektur ,,SemCrypt''.

4. Konklusion

Pseudonymisierung ist eine viel versprechende Technik, um den Anforderungen der Datenspeicherung an die Primärnutzung zu entsprechen und die Wahrung der Privatsphäre bei Sekundärnutzung zu garantieren. Der Pseudonymisierungsansatz PIPE wird derzeit in einem Unternehmen mittlerer Größe für die Speicherung von Ergebnissen prädiktiver genetischer Tests verwendet. Das Unternehmen verfolgt mit dem Prototyp mehrere Ziele: (i) Um den rechtlichen Anforderungen zu genügen, müssen genetische Daten in einem anonymisierten Zustand gelagert werden, (ii) Vertraulichkeit und Datenschutz müssen auch gegen interne Angreifer gewährleistet werden, (iii) Zugriff auf gesicherte Daten sollte für beauftragte externe Personen möglich sein, (iv) Sekundärnutzung für interne Statistiken und Forschungsprojekte muss unterstützt werden. Die Korrektheit von PIPE wurde formal mit dem Avispa-Tool (Automatisierte Validierung von Internet Security-Protokollen und -Anwendungen[2]) verifiziert.

Pseudonymisierung erfordert eine ausreichend große Zahl von Einzelpersonen und Daten, um effektiv zu sein. Wir betonen auch die Tatsache, dass eine erfolgreiche Pseudonymisierung (sowie Anonymisierung) eine zuverlässige Depersonalisierung der Basisdaten erfordert. Dies kann für bestimmte Arten von medizinischen Daten schwierig oder sogar unmöglich sein. Vor allem Daten, die genetische Informationen enthalten, müssen mit besonderer Sorgfalt behandelt werden, da diese Daten an sich identifizierende Informationen darstellen. In Kombination mit anderen, öffentlich zugänglichen Datenquellen könnten Profile erstellt werden, die zum ,,gläsernen Patienten'' führen.

Danksagung

Diese Arbeit wurde vom Bundesministerium für Wirtschaft, Familie und Jugend (BMWFJ) und der Stadt Wien im Rahmen des Kompetenzzentrums Secure Business Austria sowie der FIT-IT-Forschungsinitiative Trust in IT Systems (Fördervertrag 816158) gefördert.

Literatur

Aggarwal, C. (2005): On k-anonymity and the curse of dimensionality. In: Proc. of the 31st Int. Conf. on Very Large Databases (VLDB).

Charles, N. (2001): Telling them no. People 56 (2): 81.

Chaudry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., Morton, S. C., Shekelle, P. G. (2006): Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Annals of Internal Medicine 144 (10): 742–752.

Coalition for Genetic Fairness (2004): Faces of genetic discrimination – how genetic discrimination affects real people.

Congress of the United States of America (2008): Genetic information nondiscrimination act.

Council for Responsible Genetics: Genetic discrimination, http://www.councilforresponsiblegenetics.org/(January 2001).

European Union (1995): 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. Official Journal of the European Communities L 281: 31–50.

Evans, R. S., Pestotnik, S. L., Classen, D. C., Bass, S. B., Burke, J. P. (1992): Prevention of adverse drug events through computerized surveillance. In: Proc. of the Annual Symp. on Computer Application in Medical Care 1992: 437–441.

Fischer-Hübner, S. (2001): IT-Security and Privacy: Design and Use of Privacy-Enhancing Security Mechanisms. Berlin: Springer.

Halbert, T., Ingulli, E. (2008): Law and ethics in the business environment. South-Western College/West, 6th edition.

Kaushal, R., Jha, A. K., Franz, C., Glaser, J., Shetty, K. D., Jaggi, T., Middleton, B., Kuperman, G. J., Khorasani, R., Tanasijevic, M., Bates, D. W. (2006): Return on investment for a computerized physician order entry system. Journal of the American Medical Informatics Association 13: 261–266.

Roses, A. D. (2000): Pharmacogenetics and the practice of medicine. Nature 405: 857–865.

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., Detmer, D. E. (2007): Toward a national framework for the secondary use of health data: an American medical informatics association white paper. Journal of the American Medical Informatics Association 14: 1–9.

Thomson, D., Bzdel. L., Golden-Biddle, K., Reay, T., Estabrooks, C. A. (2005): Central questions of anonymization: a case study of secondary use of qualitative data. Forum Qualitative Social Research 6: 29.

United States Department of Health & Human Service (2006): Health Insurance Portability and Accountability Act of 1996. Public Law 104–191.

''Verraten und verkauft'' – Das Geschäft mit unseren Daten. Stern Nr. 36/2008.

Autoren

Thomas Neubauer
ist Senior Researcher am Institut für Softwaretechnik und Interaktive Systeme an der Technischen Universität Wien. Er ist außerdem Leiter des Bereichs E-Health bei ESQH Österreich. Seine Forschungsschwerpunkte liegen im Bereich Risikomanagement, E-Health und insbesondere Datenschutz. Er publizierte über 50 Beiträge auf internationalen Konferenzen und in Journalen sowie ein Patent. Dr. Neubauer arbeitete zwei Jahre im Finanzbereich und war Berater für das Österreichische Bundeskanzleramt sowie Österreichische Sozialversicherungen.

Johannes Heurix

ist wissenschaftlicher Mitarbeiter bei Secure Business Austria Research und am Institut für Softwaretechnik und Interaktive Systeme an der Technischen Universität Wien. Seit dem Abschluss seines Studiums der Wirtschaftsinformatik an der Technischen Universität Wien widmet er sich der Forschung im Bereich der Erhaltung und Verbesserung der Privatsphäre von Patienten im Bereich E-Health, insbesondere durch Pseudonymisierungstechniken. Zu seinen weiteren Forschungsinteressen gehören die Sicherheit und Integrität von Daten in öffentlichen Datenbanken (untrusted databases), auf Smartcards basierende Authentifizierung sowie die sichere Archivierung von medizinischen Akten.

[2] http://www.avispa-project.org/.

A Min Tjoa
ist Vorstand des Instituts für Softwaretechnik und Interaktive Systeme an der Technischen Universität Wien. Seine Forschungsschwerpunkte liegen in den Bereichen Data Warehousing, Data Mining 'Software Engineering' Object Oriented Analysis and Design, Semantic Web sowie Informationssysteme für Personen mit besonderen Bedürfnissen. Prof. Dr. Tjoa ist Autor/Herausgeber von 15 Büchern und über 150 Fachbeiträgen und Träger zahlreicher Auszeichnungen.

Edgar R. Weippl
ist Wissenschaftlicher Leiter des Forschungszentrums Secure Business Austria (www.sba-research.org). Seine Forschungsarbeit konzentriert sich auf angewandte IT-Sicherheit und E-Learning. Er publizierte ca. 80 Artikel in Journals und auf Konferenzen. Von 2002 bis 2004 arbeitete er als Berater für eine Health maintenance organization (Empire BlueCross BlueShield) in New York und für die Deutsche Bank in Frankfurt. 2006 gründete er gemeinsam mit Kollegen das Forschungszentrum SBA Research und habilitierte sich im Sommer 2009.