INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 80 (2011) 190–204

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

A methodology for the pseudonymization of medical data

Thomas Neubauer[a,*], *Johannes Heurix*[b]

[a] *Institute of Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstrasse 9-11, 1040 Vienna, Austria*
[b] *SBA Research, Favoritenstrasse 16, 1040 Vienna, Austria*

ARTICLE INFO

Article history:
Received 6 July 2010
Received in revised form
18 October 2010
Accepted 19 October 2010

Keywords:
Computer security
Information management
Electronic medical record
Privacy
Pseudonymization

ABSTRACT

Purpose: E-health enables the sharing of patient-related data whenever and wherever necessary. Electronic health records (EHRs) promise to improve communication between health care providers, thus leading to better quality of patients' treatment and reduced costs. However, as highly sensitive patient information provides a promising goal for attackers and is also frequently demanded by insurance companies and employers, there is increasing social and political pressure regarding the prevention of health data misuse. This work addresses this problem and introduces a methodology that protects health records from unauthorized access and lets the patient as data owner decide who the authorized persons are, i.e., who the patient discloses her health information to. Therefore, the methodology prevents data disclosure that negatively influences the patient's life (e.g., by being denied health insurance or employment).

Methods: This research uses a combination of conceptual-analytical, artifact-building and artifact-evaluating research approaches. The article starts with a detailed exploration of existing privacy protection mechanisms, such as encryption, anonymization and pseudonymization, by comparing and analyzing related work (conceptual-analytical approach). Based on these results and the identified shortcomings, a pseudonymization methodology is defined and evaluated by means of a threat analysis. Finally, the research results are validated with the design and implementation of a prototype (artifact building and artifact evaluation).

Results: This paper presents a new methodology for the pseudonymization of medical data that stores health data decoupled from the corresponding patient-identifying information, allowing privacy-preserving secondary use of the health records in clinical studies without additional anonymization steps. In contrast to clinical studies, where it is not necessary to identify the individual participants, insurance companies and employers are interested in the health status of individuals such as potential insurance or job applicants. In this case, pseudonymized records are practically useless for these parties as the patient controls who is able to reestablish the link between health records and patient for primary use – usually only trusted health care providers.

Conclusions: The framework provides health care providers with a unique solution that guarantees data privacy (e.g., according to HIPAA) and allows primary and secondary use of the data at the same time. The security analysis showed that the methodology is secure and protected against common intruder scenarios.

* *Corresponding author*. Tel.: +43 1 58801 18801.
E-mail address: neubauer@ifs.tuwien.ac.at (T. Neubauer).
1386-5056/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2010.10.016

1. Introduction

In today's health care system, the availability of reliable information has a tremendous impact on decisions regarding the patients' care and, as a result, on the quality of treatment and patients' health. Over the past years, electronic health records (EHRs) have been introduced as a method for improving communication between health care providers and access to data and documentation, potentially leading to better clinical and service quality (cf. [1]). The EHR promises the reduction of adverse drug events, which are estimated to account for about $175 billion a year in the US [2], and a reduction of the very high number of more than 150,000 cases of deaths related to adverse drug reactions each year in the US [2] as it provides physicians and their health care teams with decision support systems and guidelines for drug interactions. The EHR could achieve massive savings with the digitizing of the results of diagnostic tests and images. A study by the Rand Corporation found that adopting the EHR could result in more than $81 billion in annual savings in the US if 90% of the health care providers used it [2]. However, the electronic storage of health data raises considerable privacy concerns. In fact, the discussion of privacy is one of the fundamental issues in health care today and is often seen as a trade-off between the patient's requirement for privacy and the society's needs for improving efficiency and reducing costs in the health care system. With informative and interconnected health-related data comes highly sensitive and personal information. Due to the high sensitivity of the data, there is increasing social and political pressure to prevent the misuse of health data. It is the fundamental right of every citizen to demand privacy, because the disclosure of medical data can cause serious problems for the patient. Insurance companies or employers could use the information to deny health coverage or employment. The disclosure of sensitive data, such as a history of substance abuse or HIV infection, could result in discrimination or harassment. In addition to social and political pressure, legal acts demand the protection of health data. The Health Insurance Portability and Accountability Act (HIPAA) [3] demands the protection of patients' data that is shared from its original source of collection. In the EU the processing and movement of personal data has been legally regulated by the EU with Directive 95/46/EC [4]. A citizen's right to privacy is also recognized in Article 8 [5] of the European Convention for the Protection of Human Rights and Fundamental Freedoms. In order to protect patients' privacy when using, transferring and storing medical records, a variety of privacy enhancing technologies (cf. [6] for a definition) have been proposed. However, existing approaches often (i) do not comply with the current legal requirements (cf. [4,7,8,9,10]), (ii) do not fulfill basic security requirements (cf. [11,12,13]), and (iii) are not suitable for use with clinical studies (cf. Section 2). This work presents the pseudonymization methodology PIPE (Pseudonymization of Information for Privacy in e-Health). PIPE is used for decoupling the medical data from the patient-identifying data as well as restoring the link for authorized parties, while the actual medical records are maintained and accessed by external (health) applications. The pseudonymization methodology is based on cryptographic operations and, therefore, uses a server-side hardware security module (HSM, cf. [14]), a specially protected piece of hardware, for the execution of cryptographic operations, which ensures that the encryption and decryption operations are executed within a secure environment and that no secret key is present outside the HSM in plaintext at any time. Unlike other HSM applications that rely on the device as both a specially secured environment for encryption and decryption operations and a secure keystore, in PIPE the HSM is employed as trusted cryptographic processor only.

This research uses a combination of conceptual-analytical, artifact-building and artifact-evaluating research approaches. The article starts with a detailed exploration of existing pseudonymization protection mechanisms, such as encryption, anonymization and pseudonymization, by comparing and analyzing related work (conceptual-analytical approach). Based on these results and the identified shortcomings, a pseudonymization methodology is defined and evaluated by means of a threat analysis. Finally, the research results are validated with the design and implementation of a prototype (artifact building and artifact evaluation).

2. Background

Protection of the patients' privacy can be achieved with two different techniques, anonymization and encryption, which unfortunately both suffer from major drawbacks: While anonymization – the removal of the identifier from the medical data – cannot be reversed and therefore prevents primary use of the records by health care providers who obviously need to know the corresponding patient (as a minor point, patients cannot benefit from the results gained in clinical studies because they cannot be informed about new findings), encryption of the medical records prevents them from being used for clinical research (secondary use) without the explicit permission of the patient, who has to decrypt the data and, in doing so, reveals her identity. Considering that some medical records tend to be very large (up to hundreds of MB [15]), encryption could also be a very time-consuming operation [16]. A method that resolves these issues is pseudonymization, where identification data is transformed and then replaced by a specifier that cannot be associated with the identification data without knowing a certain secret [17,12,18]. Pseudonymization allows the data to be associated with a patient only under specified and controlled circumstances. A pseudonymized database must contain at least two tables, one where all the personal data is stored, and one where the pseudonyms and the pseudonymized data are stored. The process of identifying and separating personal from other data is called depersonalization (cf. [19]). After depersonalization and subsequent pseudonymization, a direct association between individuals and their data can only be established under strictly defined circumstances.

2.1. *Pseudonymization*

However, existing pseudonymization approaches and systems have a variety of shortcomings: Thielscher et al. (cf. [20]) developed a system consisting of two databases, one for the

patient's personal identification data and the other for the medical data. While the datasets are stored in a decoupled way, the relationship between the patient and her medical records can be restored with secret keys stored on a smart card. The secret keys generate unique data identification codes, which are also stored in the database and do not contain any patient-identifying information. Authorizations are granted by sharing these codes between the patient and health care providers where these authorizations are valid for only a certain period of time. As a fallback mechanism in case a patient loses her smart card, a centralized patient-pseudonym list is maintained, as otherwise there would be no way to recover the identifier. This centralized patient-pseudonym list could be the target of intrusion attacks. Thielscher et al. circumvent this security flaw by operating the list offline. This organizational work-around seems to promise a higher level of security until an insider attack is conducted (cf. [21,22,23]) or an attacker gains physical access to the computer that holds the list.

The approaches developed by Pommerening (cf. [24,25]) are only applicable for secondary use, pseudonymizing the data for transport, and rely on a combination of hashing and encryption techniques. The different approaches apply to different scenarios: (i) Beginning with medical data from overlapping sources for one-way secondary use, a unique patient identifier is replaced by a pseudonym generated by a one-time encryption operation conducted at a pseudonymization service provided by a trusted third party. Furthermore, the medical data is encrypted with the public key of the secondary user to ensure one-time secondary use. (ii) To extend this scenario with the possibility of re-identification, the one-way pseudonymization is replaced by a reversible encryption. An important prerequisite for this scenario is that the unique patient identifier is not available publicly, but generated specifically for the project in question by a second trusted third-party service. (iii) The final model involves multiple secondary users in a research network, where a central research database is introduced containing the medical data and the unique patient identifier, again generated by a trusted third-party service. Each secondary user accesses the medical data using the pseudonymization service. The drawback of the Pommerening approaches are the trusted third party services. While they provide security against external attacks, an insider, perhaps bribed by an attacker, could still abuse the service.

El Kalam et al. [26] propose a very similar approach that involves hashing and encryption for the secondary use of pseudonymized data. In this case, a unique patient identifier, which may not be publicly known, is used for deriving an anonymized identifier. Unlike in the Pommerening approaches, this patient identifier is not generated by a trusted third party but within the secured environment of a smart card. The identifier also stays within the smart card and should never be disclosed. Furthermore, a unique project identifier is required as input, so that the anonymized identifier is the result of a one-way hashing procedure with the concatenation of the unique patient identifier stored on the smart card and the unique project identifier as input. That way, the patient has to explicitly consent to each secondary use of her data in each individual project. In order to prevent attacks where an external adversary tries to link data held by two different hospitals by knowing the fingerprint (hash value) of a certain patient/project and gaining unauthorized access to the database of the hospital where the project is conducted, the anonymized identifier is encrypted with a secret key only known to the hospital. The corresponding decryption key is only known to the project. Although this concept provides protection against external attackers, a malicious insider can still link the medical data to the corresponding patient by knowing the fingerprint. Another issue is the smart card as generator and storage provider for the unique patient key. As this key never leaves the smart card, its uniqueness cannot be ensured. Furthermore, lacking a suitable backup mechanism for the unique identifier, the medical data cannot be tracked back to the patient if she loses the smart card.

Noumeir et al. [27] describe the pseudonymization of radiology data encoded as DICOM [28] files for secondary use. The unique patient identification numbers in the DICOM images are replaced with pseudonyms and the files are stored in a separate research database. Here, the authors distinguish between two kinds of pseudonyms: irreversible one-way pseudonyms (i.e., anonymization) and reversible pseudonyms. One-way pseudonyms are generated by processing the patient identification number with a hashing algorithm. As hashing is prone to collisions, the authors propose to additionally hash the patient's medical history. In order to prevent dictionary attacks, the technique of salting, i.e., concatenating a random value to the input, can be applied. Alternatively, hash-based message authentication code techniques requiring a secret key may be used. Reversible pseudonyms also involve a secret key for the encryption of the patient identification number. Noumeir et al. propose the application of the DES algorithm using 64 bits of input, which is especially suitable for the DICOM patient identification field that is constrained to a maximum of 64 characters. Although the authors state that the secret keys need to be protected, no suggestion is made concerning how to protect the keys.

The approach developed by Peterson [29] involves the use of multiple encryption keys and three data tables to make personal medical data available without compromising the patient's privacy. During registration, the patient is issued a unique global key (GK) and a server side key (SSID). Furthermore, she has to provide a unique personal encryption key (PEK) and a password. This information is stored in the user table. The security table contains the reference to the user (SSID), a server side encryption key (SSEK) and the reference to the personal data in the personal data table. The personal data is doubly encrypted with the PEK and the SSEK. Data retrieval requires the knowledge of either the PEK or the GK (can be printed on an ID card), which is transferred to the server, which looks up the referenced SSID. With the SSID, the personal data record can be retrieved and decrypted with the corresponding keys. Data addition, deletion, and modification require the knowledge of the password in addition to the PEK or GK. As a fallback mechanism in case of a lost smart card, the patient can log in with the PEK and password and is issued a new GK, which can be printed on a new ID card, invalidating the old GK. This approach comes with some seri-

ous drawbacks: As all keys needed for decrypting the medical data are stored in the database, an attacker gaining access to the database could decrypt all information. Even more importantly, as the password and the keys are stored in the database, the attacker could change data stored in the database. Furthermore, the PEK poses a security flaw as the user is required to enter new PEKs during registration until a unique one has been found. This mechanism allows a possible attacker to immediately gain access to the medical data related to this key.

The architecture for the realization of the German Electronic Health Card (EHC) proposed by the Fraunhofer Institute for Software and System Engineering [30,31] and supported by the German Federal Ministry of Health is designed as a service-oriented architecture divided into five layers realizing different service applications, such as emergency data, electronic prescription, or electronic health record retrieval. Access to the medical data stored in a virtual file system is handled with the ticket toolkit concept involving hybrid encryption: The medical data is encrypted with a unique session key, and this key, in turn, is encrypted with the public key of the authorized user. The asymmetric keypair is stored on the health card. The ticket toolkit contains a ticket builder, a ticket verifier, access policy information and an encrypted link to the file (or directory) in the virtual file system. Each file and directory is assigned a default ticket toolkit and a number of private ticket toolkits defined by the patient for authorized users. If no private ticket toolkit is available for the user requesting a certain file, the system returns the default ticket toolkit based on a challenge that the data requestor has to solve in order to gain access rights. As the asymmetric keypair is stored on the health card, which is prone to loss or damage, a second private ticket toolkit is generated for each entry relying on a second asymmetric keypair stored on an emergency card (e.g., a relative's card). This second keypair can then be used to decrypt the session keys in order to re-key them with the new public key stored on the new health card. While the ticket concept ensures confidentiality, fully encrypting the medical data is time-consuming, especially for large medical images, as well as unsuitable for secondary use.

Stingl et al. [32] also rely on the encryption of medical data in order to realize a centralized and secure health data repository. Similarly to the EHC concept, the actual medical data is encrypted with a symmetric key and the patient shares this key with her trusted associates, encrypted with the public key of their individual asymmetric keypair representing their data access authorizations as clearance packages. The documents are stored in a hierarchical manner and the data repository is pseudonymized by obscuring the master/detail relationship by encryption. Thus, only those who are in possession of the correct key are able to establish the link between patient and health data. The scheme also specifically distinguishes between data creator, owner, access grantor, and grantee. Like the EHC, this approach suffers from the problem of fully encrypted medical data. In [33,34], Stingl and Slamanig adapt this concept to be used in a pseudonymized e-health portal where the same roles are applied and the medical data is encrypted in a hybrid manner. In addition, the e-health portal concept introduces an identity management system where each user is in possession of multiple sub-identities, one publicly known and several private sub-identities. Each sub-identity is related to an asymmetric keypair stored on a smart card and accessed by individual PINs. The public sub-identity is used for data access authorizations (involving the public sub-identity of the access grantor and grantee, creator, and patient) where the public sub-identity of the receiver is then replaced by a private one. The publicly known sub-identity and the corresponding public key, however, open the architecture to impersonation attacks, e.g., for sending fake medical documents.

2.2. *Hardware security module*

The cryptographic operations required in the proposed pseudonymization procedures are best provided by a dedicated and secure hardware-based cryptographic system, i.e., a hardware security module (HSM) [14]. An HSM is an encapsulated and tamper-resistant hardware module that is designed to withstand logical as well as physical attacks [35,36]. The protection measures range from solid metal casings to special switches that zeroize the memory when tampering is detected [37]. HSMs provide standardized interfaces to communicate with their host computers such as PKCS#11 [38] and serve as secure keystore and cryptographic processors, often also providing secure key generation. HSMs are deployed in multiple application areas where security is of utmost importance: Anderson et al. [37] identify the application of HSMs in automated teller machines for PIN management including PIN acquisition/verification/generation, in electronic payment schemes as an integral part of the back-end systems at banks processing the transactions, or in military applications as encryption and decryption modules for highly sensitive communication or as nuclear command and control tools.

Wherry [39] recognizes the need for an HSM in public key infrastructures (PKIs) to protect the most important entities in PKIs, the cryptographic keys. Lorch et al. [40] utilize an HSM as a secured online credential repository in a grid PKI. Rössler et al. [41] apply the HSM to e-voting as an electronic ballot box. In this scheme, the voter receives a ballot and a voting token containing her voting ID after authentication with a signature card (smart card). The ballot is then encrypted and transferred to the election authority where the votes are decrypted only in the HSM during the counting procedure. The private decryption key is only available within the HSM so that the vote cannot be decrypted insecurely outside the HSM. Baldwin and Shiu extend the concept of HSMs to not only provide a secure keystore and secure cryptographic operations, but to provide complete security services, resulting in hardware security appliances (HSAs) as demonstrated in [42] to secure Web services. Another example of securing Web services is given by Mont et al. in [43], while Baldwin and Shiu apply HSAs to enhancing trust relationships in an outsourcing scenario of a merchant's online presence [44]. In [45], Ferreira et al. propose the application of an HSA in an accountability model suitable for health care environments where the HSA encapsulates the accountability service responsible for recording events and providing access for retrieving the recorded events only for authorized roles.

3. The PIPE methodology

PIPE (Pseudonymization of Information for Privacy in e-Health) is a novel protocol for the pseudonymization of health data that differs from existing approaches in its ability to securely integrate primary and secondary use of health data. PIPE provides a form of "traceable anonymity" to ensure the confidentiality of health records.

3.1. Overview

Patient-identifying details are separated from the actual health data, resulting in decoupled data records, i.e., identification and health records. The relationship between the patient represented by her identification record and her health data is established with pseudonyms (identification and health pseudonyms) whose link is identifiable only under specific conditions. The links between identification and health pseudonyms are encrypted with secret keys that can only be used by authenticated and authorized persons, who can use them to re-link the personal patient data with the health data for primary use. Pseudonyms are also used for data access permissions, i.e., defining new pseudonyms for access authorizations (shared pseudonyms). Revoking access rights is easily accomplished by deleting these pseudonyms.

Because of the decoupled storage scheme, secondary use of the medical records for research is privacy-preserving in that the secondary user is not able to trace the individual health records back to the patient in question. The same is true for any external or internal attackers (e.g., database administrators) who may gain access to the database system.

In order to authenticate registered users, each user is issued a security token containing the authentication credentials that grant access to the secret cryptographic keys. As hardware security tokens are prone to loss or damage, a backup mechanism is provided based on the principle of threshold schemes [46].

A server-side HSM acts as cryptographic module for executing the necessary cryptographic steps within a trusted secure environment. The cryptographic operations include all encryption and decryption operations required for functions such as user authorizations and authentications. The client-side cryptographic operations required for the challenge/response style authentication procedure are carried out with the user-owned security token acting as secure keystore for the authentication credentials and as trusted client-side cryptographic module. In this context, a security token is a secured (contact) micro controller smart card (cf. [47,48]) that has similar properties to an HSM, i.e., tamper resistance, secured key storage area, and dedicated hardware-based cryptographic engines for common algorithms such as RSA or 3DES. Due to these properties, the token is used as user-owned PIN-protected secure keystore for the authentication credentials as well as a trusted client-side cryptographic engine. The smart card requires a certified card reader (Common Criteria EAL3+) with an integrated keypad. The integrated keypad ensures the secured entry of the required PIN, preventing its exposure to potential malicious code installed on the host computer (malware, viruses, etc.). As an alternative to the smart card, the authentication credentials can be stored in a secured area of the user's computer's hard disk or on a comparable storage device (e.g., a USB stick), while the necessary client-side cryptographic operations during authentication need to be executed by a client application. In contrast to the smart card solution, this alternative is less secure due to the software-based implementation of the cryptographic algorithm.

Aimed to provide a pseudonymization service, PIPE can be applied to different scenarios (cf. Fig. 1): In the local scenario, the PIPE server pseudonymizes only records stored in the local (health) data repository and makes them available to a local (health care provider's) workstation where both patient and health care provider interact with the pseudonymization server as part of a health care provider environment (e.g., with a hospital information system). In an alternative central scenario, the PIPE pseudonymization server is responsible for providing linking information to different health records stored at distributed locations. In Fig. 1, two separate health care provider environments exist where the individual workstations have direct access to their local data repositories. Via the pseudonymization service, the health care providers are able to access records of other domains if they are explicitly authorized to do so. In this scenario, the patient also has the opportunity to retrieve the records at home.

3.2. Security model

The PIPE protocol uses a combination of symmetric and asymmetric cryptographic keys to realize a logical multi-tier hull model with three different layers (cf. Fig. 2), where each layer is responsible for one step in the data access process. The user has to pass all layers in order to retrieve the actual health data records.

The outer public and outer private keys form the outer layer, the *authentication layer*, which is responsible for unambiguously identifying the corresponding user. Together with the user's identifier, the outer private key represents the authentication credentials, which are stored along with the server's public key on the user's smart card. In combination with the correct PIN, the smart card provides two-factor authentication, where the authentication procedure involves both the user's and the PIPE server's outer keypair, the user's identifier, and two randomly selected challenges. The middle layer, the *authorization layer*, consists of the user's inner asymmetric keypair and the inner symmetric key. While the user's outer private key is created on the smart card when the card is issued to the user and never actually leaves the card, the other keys are stored in the pseudonymization database where the secret keys are stored encrypted: the inner symmetric key is encrypted with the inner public key, while the inner private key is encrypted with the outer public key. Thus, the outer private key is required to decrypt the inner symmetric key, which finally decrypts the pseudonyms and thus provides access to the innermost layer, the *pseudonymized data layer*. Encrypting the pseudonyms with a secret symmetric key instead of the inner public key is more powerful and also prevents chosen-plaintext

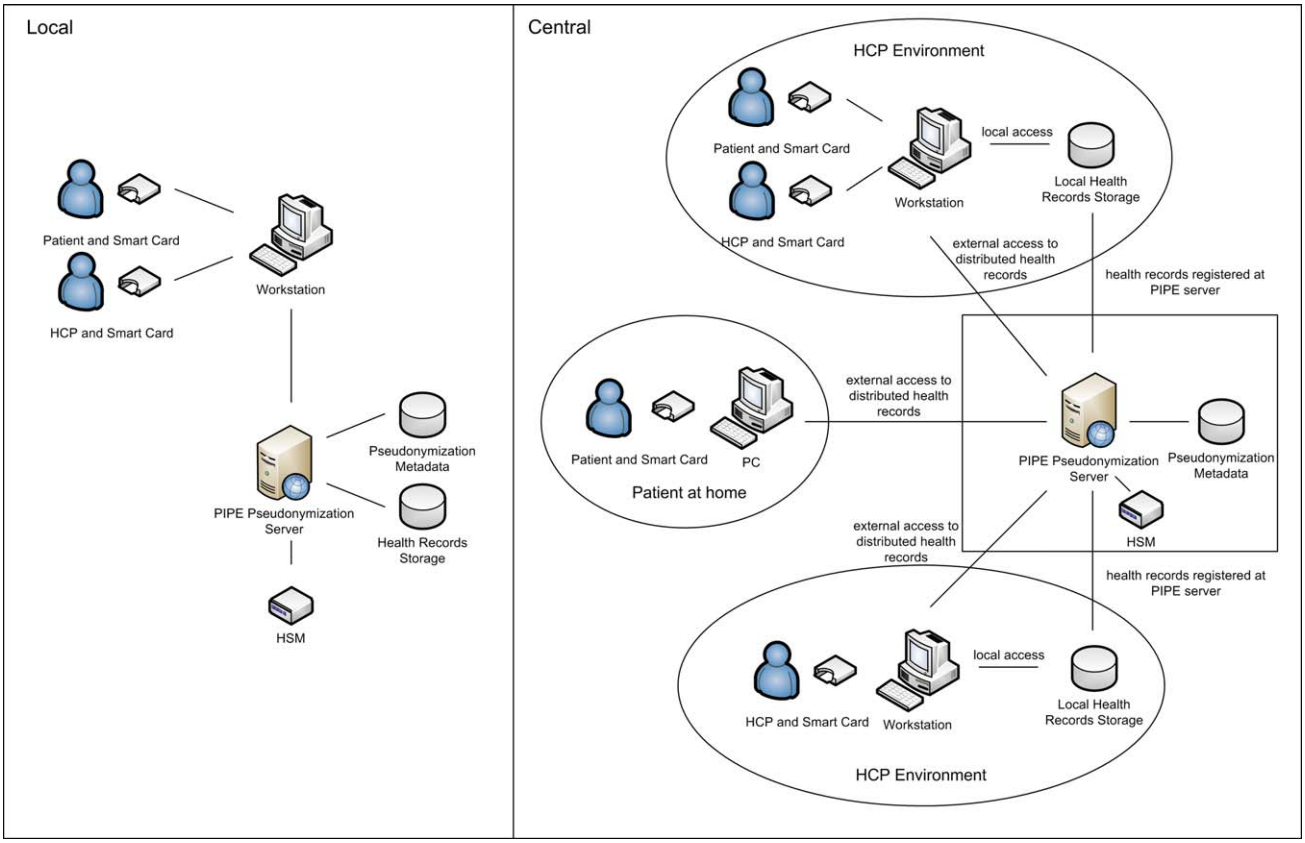

Fig. 1 – PIPE application scenarios.

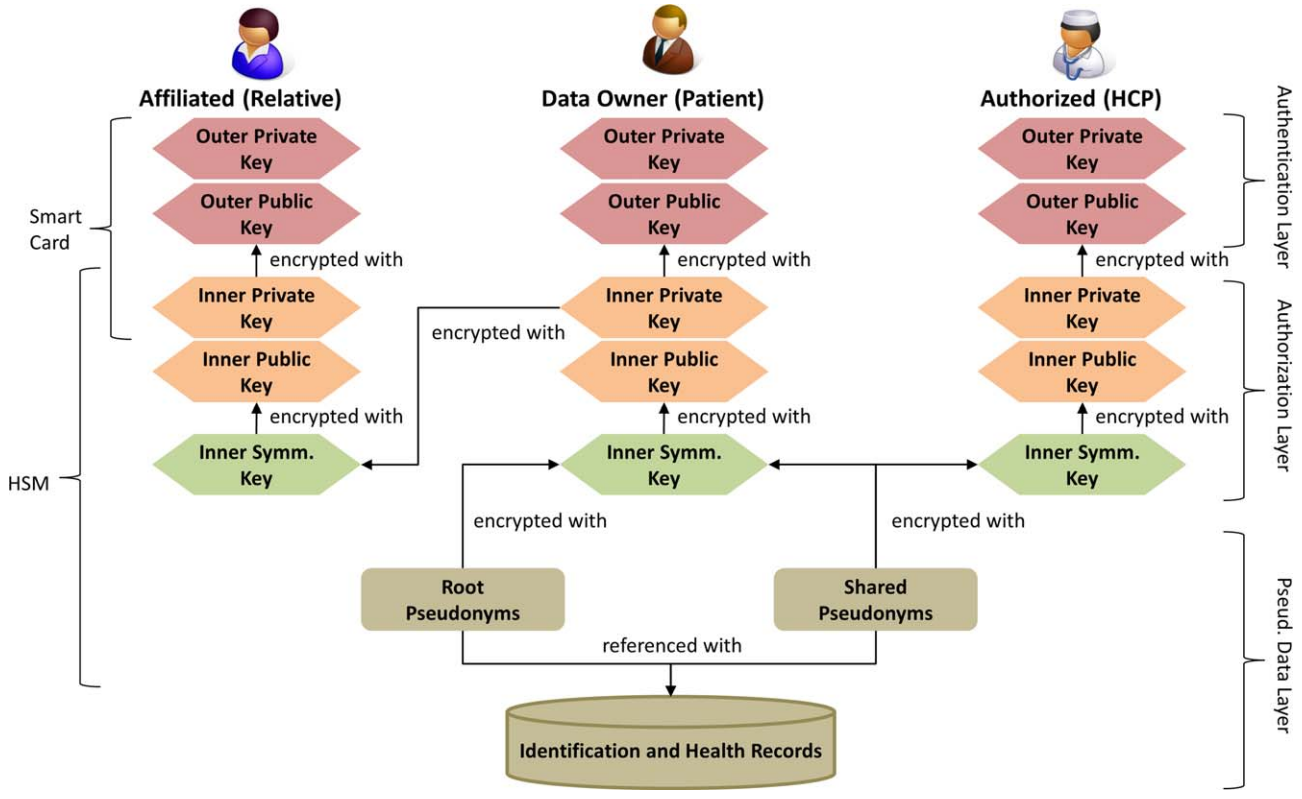

Fig. 2 – Layer-based security model.

attacks when using deterministic asymmetric cryptographic algorithms.[1]

The use of HSMs and smart cards as cryptographic modules works as follows: While the cryptographic functions of the smart card are required only for client-side authentication steps and encryptions/decryptions involving the session key used for protecting the communication channel between user/workstation and pseudonymization server, the HSM handles the major cryptographic tasks, including the server part of the user authentication, encryption/decryption of the pseudonyms, the decryption of the user's inner symmetric key with her inner private key, and the calculation and encryption of the backup shares. These shares are created for securely distributing parts of the user's inner private key to dedicated share holders (operators, see below) so that it can be reconstructed should the smart card be unavailable (lost or damaged). Storing the user's keys (with the exception of the outer private keys) in the database and loading them when required instead of keeping them in the HSM at all times makes replacing the HSM easier in the case of failure or theft. The server's outer private key still needs to be backed up by, e.g., threshold secret sharing similar to backing up the user's inner private keys.

As depicted in Fig. 2, we support three different main roles, the data owner, the affiliated, and the authorized. The data owner is represented by the patient who is in full control of her health data in that she can create data access authorizations for specific health records (for authorized users) as well as granting full access rights equivalent to root access (for affiliated users). Authorizations in this context do not refer to access authorizations in the traditional sense but rather to providing the authorization grantee with the ability to re-link identification with health pseudonyms, thus allowing them to reconnect the patient with the corresponding health records. The data owner relies on pseudonyms which are initially[2] known only to her, the root pseudonyms, as her primary health record retrieval means.

An authorized user, e.g., a trusted health care provider, is provided with the knowledge of the link between a particular health record and the corresponding identification record. This is realized by introducing a new set of pseudonyms, the so-called shared pseudonyms. As their name implies, the shared pseudonyms are shared between the data owner and the authorized person and are encrypted with both their inner symmetric keys so that both are able to decrypt this authorization relation. For each individual authorization, i.e., for each authorized user/health record combination, a new shared pseudonym is created, ensuring unobservability of access rights.

In contrast to authorized users, an affiliated user, e.g., a close relative, is entrusted with the data owner's inner private key and is therefore able to decrypt the data owner's inner symmetric key, granting the affiliated user full access to all data corresponding to the data owner. Therefore, the affiliated user is able to decrypt the links between all root and shared pseudonyms related to the data owner. The conceptual data model is depicted in Fig. 3.

The identification and health pseudonyms always form a 1:1 relationship and are referenced with their corresponding document type where this reference is stored in cleartext (record/pseudonym mapping). The link between the identification and health pseudonyms is stored encrypted with the user's inner symmetric key (pseudonym/pseudonym mapping): while the root pseudonyms are encrypted with the data owner's (patient's) inner symmetric key only, the shared pseudonyms are encrypted with both the data owner's and the authorized user's (health professional's) inner symmetric key so that both users are able to decrypt them using their corresponding ciphertexts. The link between the identification and health record is hidden and represented by the link between identification and health pseudonyms. Each health record is assigned exactly one root health pseudonym while each identification record has multiple root pseudonyms, depending on the number of health records, due to the 1:1 relationship. The health record is assigned a number of shared health pseudonyms according to the number of individual authorizations for that particular health record.

To query for particular pseudonyms (and thus records), each health pseudonym is assigned a particular keyword (the same keyword is assigned to the root and health pseudonyms that are referenced with the same health record), which can also be used for different health pseudonyms/records if applicable. To prevent the processing overhead of querying within encrypted data, we propose to store the keywords in cleartext to be usable/shared by all users, while the link between the health pseudonyms and keywords is realized by the keyword identifier. Again, the link (keyword identifier/pseudonym mapping) is encrypted with the owner's and/or authorized user's inner symmetric keys. As arbitrary keywords are ill-suited for range queries and may include confidential or patient-identifying information, the keywords are highly structured and constructed from pre-specified keyword templates: document type (image, text, etc.), disease type (e.g., International Statistical Classification of Diseases and Related Health Problems (ICD) [49]), and date.

Note that depersonalization (i.e., decoupling identification data from health data records) is an important factor to allow the effective pseudonymization of medical data. In certain cases it is not sufficient to just remove the patient's name from the medical record to depersonalize them. Some medical areas (e.g., genomics) require more in-depth inspection, because genome sequences, which have identifying properties per se, may require splitting up the records for effective depersonalization. However, in-depth discussion of this topic is outside the scope of this work. For this paper we assume that the medical data records have already been depersonalized.

4. Workflows

In this section, we describe the main PIPE workflows including authentication, user authorization, user affiliation, health

[1] Pseudonyms are stored in cleartext when mapped to a particular record while the link between them is hidden by storing the pseudonyms encrypted in a single relation.

[2] By affiliations via key-sharing, the affiliated user is granted access to the root pseudonyms as well.

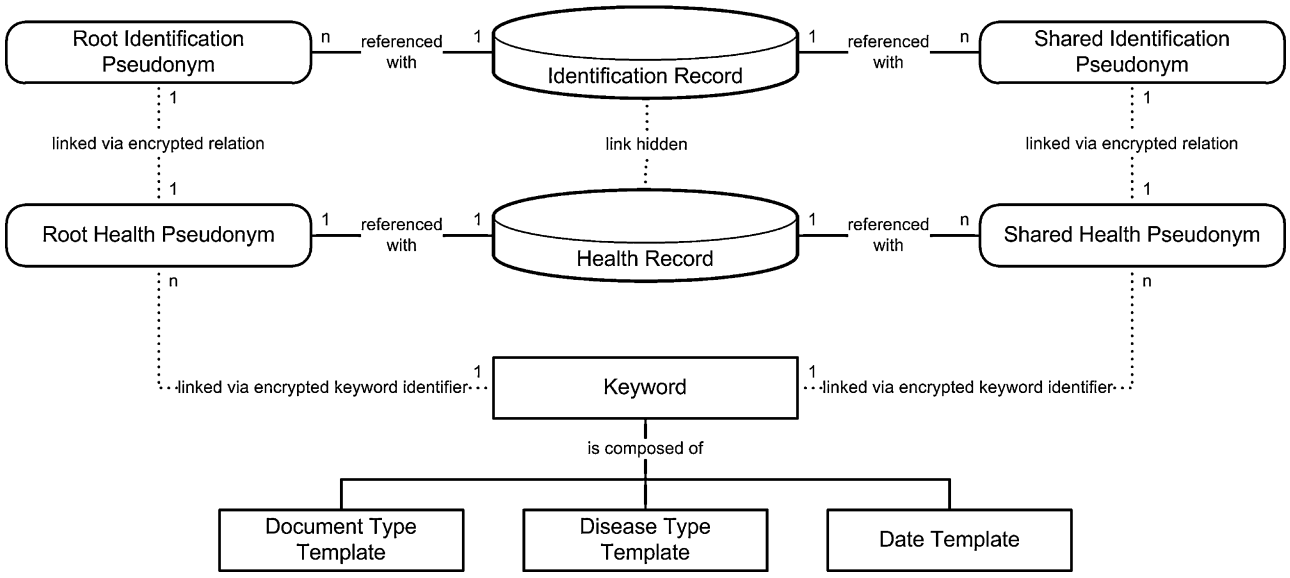

Fig. 3 – Data model.

Table 1 – Abbreviations.

OPuK	Outer public key
OPK	Outer private key
IPuK	Inner public key
IPK	Inner private key
SK	Session key
ISK	Inner symmetric key
ID	Identification (pseudonym or record)
HE	Health (pseudonym or record)
u	User
ow	Data owner
au	Authorized user
af	Affiliated user

data storage, and retrieval. Abbreviations used in the workflow diagrams are summarized in Table 1.

Session key encryption/decryption is actually needed only when transferring data over an untrusted network. Alternatively, established security mechanisms such as TLS[3] can be applied. In the local scenario with trusted communication channels, SK encryption is not required and may involve unnecessary cryptographic overhead, especially when records are fully encrypted for transport. For the sake of completeness, all workflow descriptions in this section involve the SK.

4.1. *User authentication*

User authentication (Fig. 4) involves the mutual authentication of the user using the smart card and the server, involving their outer keypair and two nonces (randomly selected numbers used once) as user/server challenges. Once both identities are confirmed, the user's inner private key is retrieved from the pseudonymization database and transferred to the user's smart card to be decrypted with the user's outer private key. With the decrypted inner private key, the user's inner symmetric key can be decrypted within the HSM at the pseudonymization server and be cached for further operations along with the user's inner private key. In addition, a session key is generated at the HSM and securely (via encryption) transported to the user's smart card so that the key appears in cleartext only on the smart card and HSM.

4.2. *Health data retrieval*

To retrieve a particular health record (Fig. 5), the user first needs to query for the particular encrypted pseudonyms by creating a keyword using the keyword templates, retrieving the corresponding keyword identifier, and querying for the encrypted identifier to find matching encrypted pseudonyms, i.e., the encrypted pseudonym mappings associated with the encrypted keyword identifier. The pseudonym pairs are then decrypted with the user's inner symmetric key and the plaintext pseudonyms then used to retrieve the corresponding identification and health records, which are transferred to the user to be displayed (possibly merged). Optionally, the pseudonyms and keyword identifier are also transferred to the user (root pseudonyms for authorizations). The record retrieval procedure is the same for the patient as data owner, health care provider as authorized user, and relative as affiliated user, with the difference that the patient and relative both query for the patient's root pseudonyms, while the health care provider relies on the shared pseudonyms.

4.3. *User authorization*

To provide a trusted health care provider with the knowledge of the link between the patient's identification record and a particular health record (Fig. 6), a new shared pseudonym pair is created as authorization relation. The patient first has to

[3] Transport Layer Security.

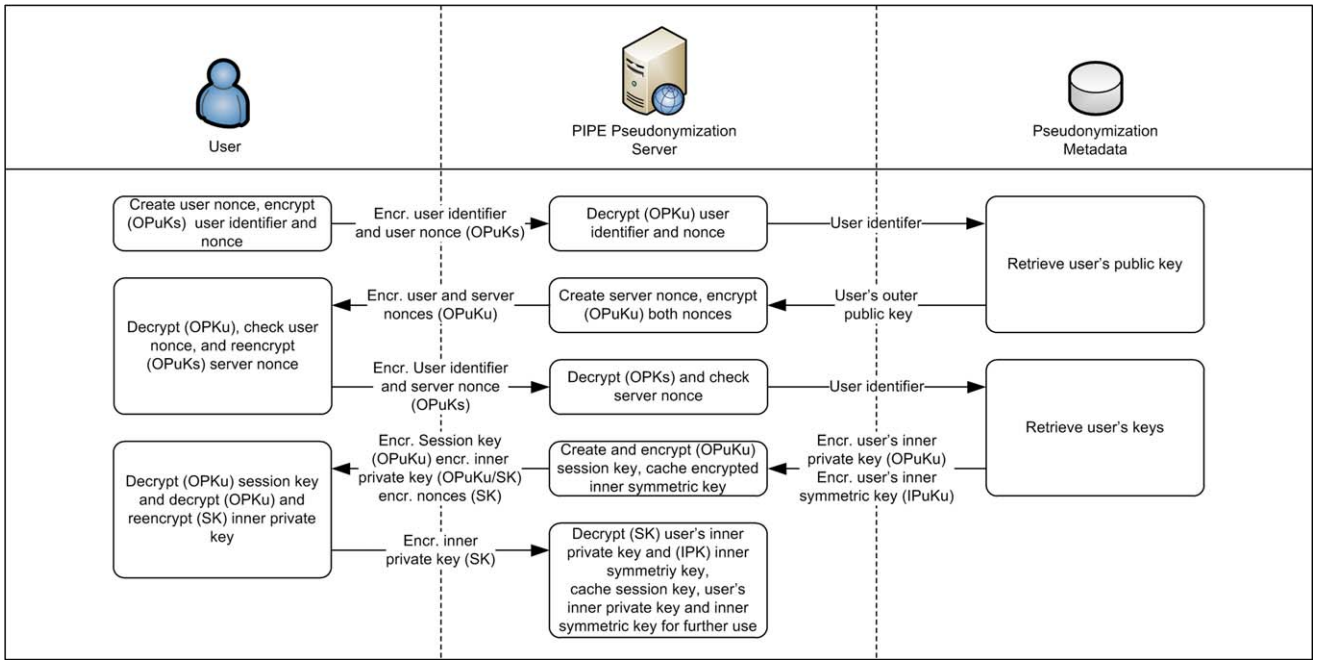

Fig. 4 – User authentication.

retrieve the root pseudonym pair and keyword identifier corresponding to the health record he or she intends to share with the health care provider. Furthermore, both the patient as data owner and the health care provider as authorized user have to be authenticated at the same workstation so that both user identifiers are available at the client side, while both inner symmetric keys are cached at the HSM of the pseudonymization server.

The root pseudonym pair is then transferred to the pseudonymization server along with both user identifiers and the keyword identifier, and the corresponding record identifiers retrieved using the cleartext record/pseudonym mappings.

The server then randomly selects a new shared pseudonym pair, which is first encrypted with both users' inner symmetric keys (along with both identifiers and the keyword identifier) and then stores them in the database as authorization relation. Finally, the cleartext pseudonyms are then referenced with the retrieved record identifiers to create two new record/pseudonym mappings.

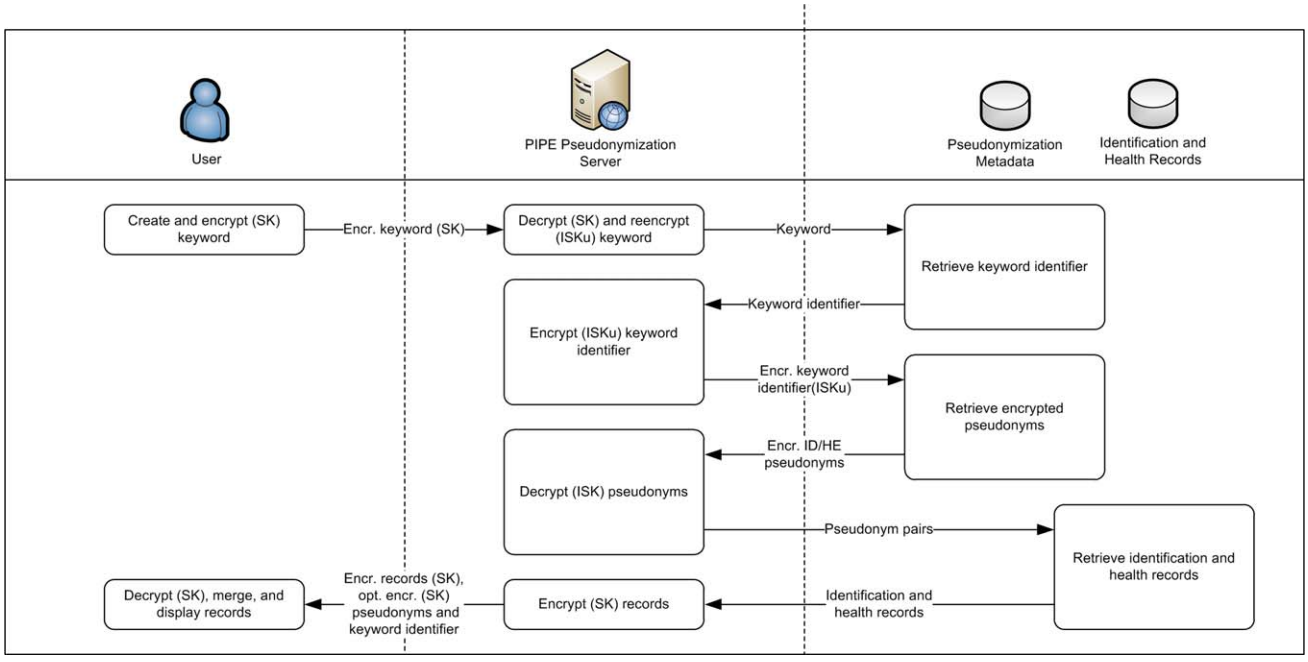

Fig. 5 – Health data retrieval.

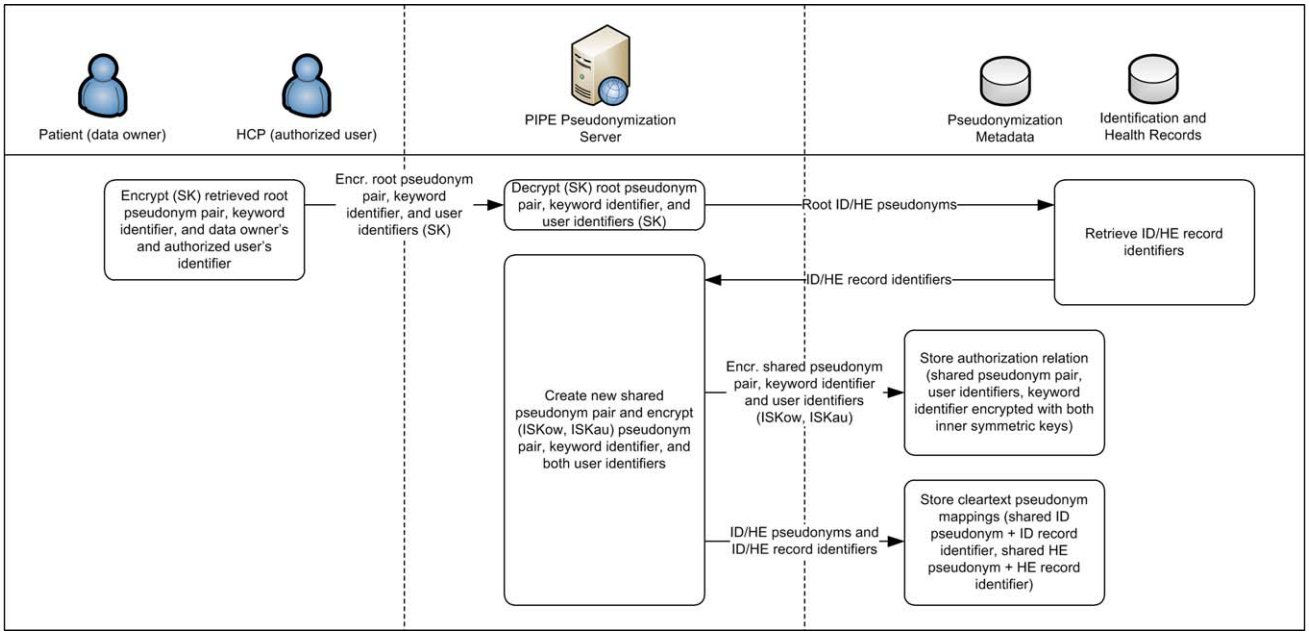

Fig. 6 – User authorization.

4.4. *User affiliation*

As with authorizations, a user affiliation (Fig. 7) requires that both the patient as data owner and the trusted relative as affiliated user are authenticated at the same workstation. Then both user identifiers are transferred to the pseudonymization server where they are encrypted with both users' inner symmetric keys. In addition, the patient's inner private key is also encrypted with the relative's inner symmetric key, and all elements are stored in the pseudonymization metadata storage as affiliation relation.

4.5. *Health data storage*

From the viewpoint of the patient as data owner, health data storage (Fig. 8) first requires that an 'old' root identification pseudonym is retrieved as reference to the identification record. Furthermore, the patient creates a new keyword and enters the new health record into the workstation. Then the pseudonym, new keyword, new health record, and user identifier are transferred to the pseudonymization server, where the keyword is stored (and its identifier determined by the database engine) and the identification record identifier retrieved. The new record is stored in the health records database and its record identifier returned to the server. Then, the server creates a new root pseudonym pair and stores it encrypted with the keyword identifier and user identifier as root access, as well as the cleartext record/pseudonym mappings.

4.6. *Health professional as health record provider*

While in the previous section the patient is described as data provider, in most cases, the main health care provider is actually a trusted health professional. The PIPE framework requires a particular data adding authorization where the health care provider is specifically authorized in advance to add new records for the patient. A data adding authorization is basically a combination of user authorization and health data storage in that both a new root and shared pseudonym pair is created and stored as authorization and root access relations. Instead of referencing the health pseudonyms with an exist-

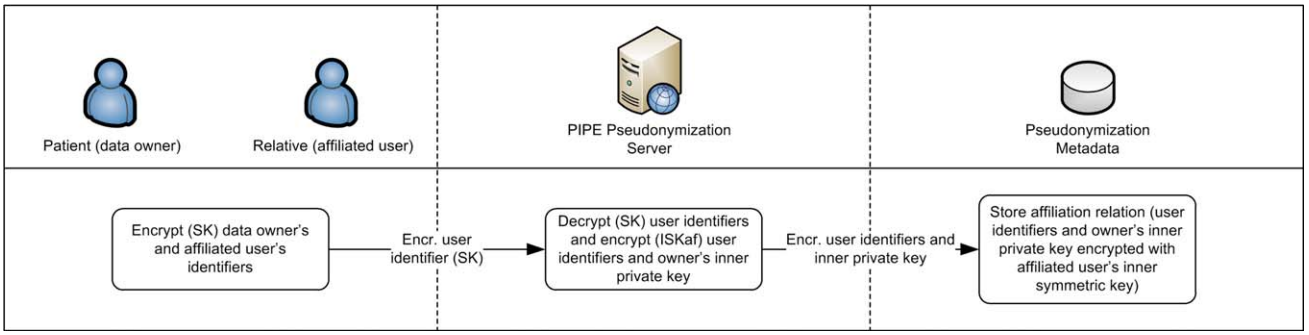

Fig. 7 – User affiliation.

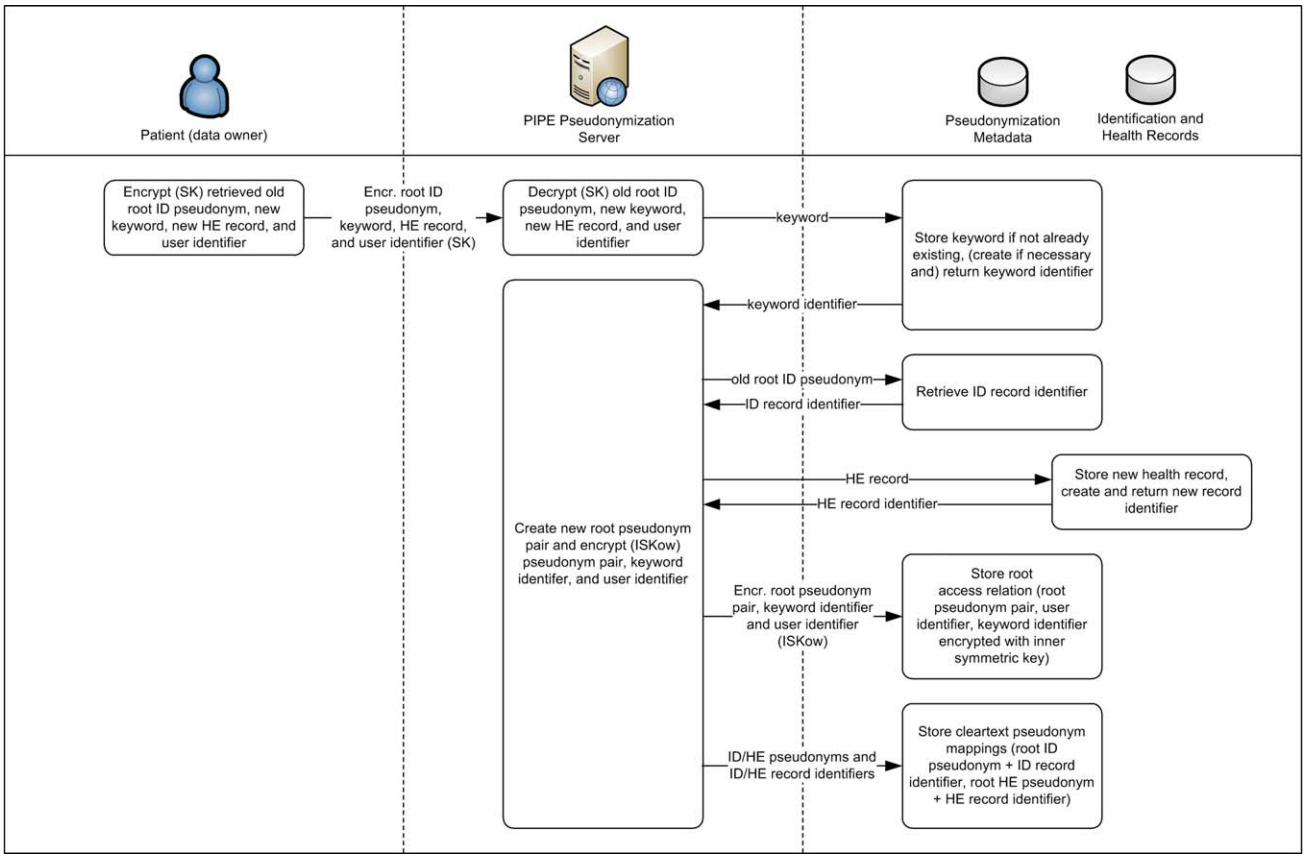

Fig. 8 – Health data storage.

ing health record, they are assigned to a placeholder that is stored instead. When the new health record becomes available, the health care provider needs to query for the particular shared pseudonym (via keyword) to replace the placeholder with the actual new health record.

5. Prototype

As a proof-of-concept, we implemented a prototype for the Microsoft® Windows platform using .NET technology. As cryptographic algorithms, RSA with 2048 bit keys and AES with 256 bit keys are used and pseudonymization metadata is stored in a relational database (MS SQL Server). This prototype realizes the core functionality of user management, record storage and retrieval, and authorization and de-authorizations, but also includes several advanced functions such as asynchronous authorizations and a notification system. The prototype was implemented with the following requirements in mind:

- Storage of health data in an anonymized state to comply with legal requirements.
- Confidentiality/privacy protection against attackers (especially insiders).
- Data access for selected authorized personnel (external employees).
- Anonymized secondary use for internal statistics and research studies.

Our case study is established in the area of predictive genetic testing. In this scenario, the data owner is represented by the patient who orders a particular test concerning the health issues she is interested in (e.g., cardiovascular diseases) and stores the medical history. Depending on the selected test type, different genetic loci of the patient's genetic sample are analyzed for single nucleotide polymorphisms (SNP), which are then recorded by a lab technician in a lab result record. This record containing SNP compositions is then made available to a domain specialist via authorization (possibly multiple specialists depending on the domain area(s) that need to be covered), who provides the medical interpretation in the form of a medical report. The complete set of records is then made accessible to the trusted general practitioner who works out an individual health plan. This scenario covers all pseudonymization-related operations including data storage (patient, lab technician, domain specialist), retrieval (all), authorizations (lab technician, domain specialist), and affiliations (general practitioner).

The health documents include samples of medical histories, lab results, and medical reports, and were encoded in a custom HL7 CDA [50] based scheme: The header sections containing patient-related and administrative information are used as identification records[4] and the body sections contain-

[4] In our test scenario, the same header section was used for all body elements corresponding to the same patient. However, our

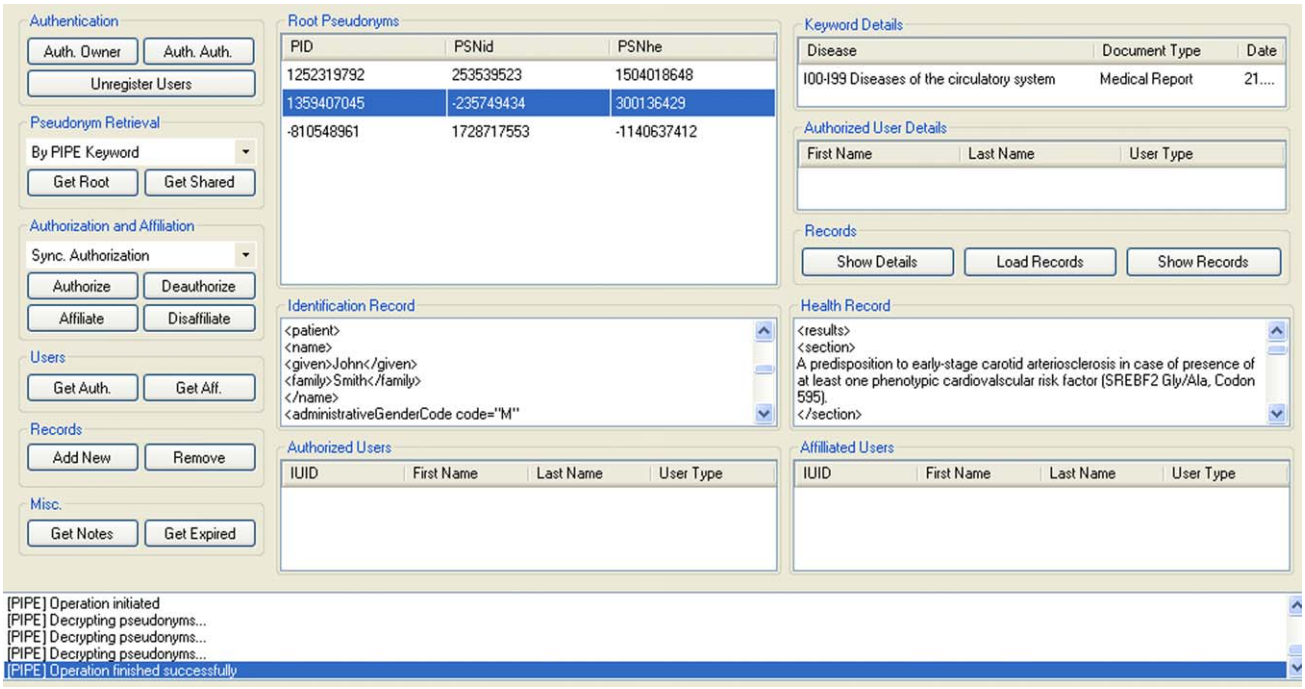

Fig. 9 – PIPE test interface with retrieved medical report.

ing the actual health data as health records. As keywords, we used ICD10-encoded disease types and simple document designators (identification, medical history, lab results, medical report).

In Fig. 9, our test interface is shown, where the query returned three root pseudonym pairs representing three documents; the currently selected pair relates to the patient John Smith and a medical report containing medical interpretations.

The prototype met the requirements stated above as follows:

- *Legal requirements*: The pseudonymized data structure contains the genetic information in an anonymized state by splitting the headers from the body sections of the documents, thus complying with legal requirements.
- *Insider attackers*: Access to the de-pseudonymized records is limited to a need-to-know basis decided by the patient, limiting the impact of potential data leakages.
- *Data access*: Valid users are issued personal security tokens that uniquely identifies and authenticate them. Record retrieval operations require authorizations resolvable by these tokens only.
- *Anonymized secondary use*: As the records are already stored in a pseudonymized state, privacy-preserving statistical inquiries are possible without further anonymization.

Performing tests involving query and retrieval operations of pseudonymized documents revealed the following results: As expected, pseudonymization resulted in a measurable impact on retrieval times due to the increased overhead of encryption/decryption and additional database operations. However, while we expected cryptography to have a major impact, the main performance hit was caused by the additional database lookups, especially that of retrieving the encrypted pseudonyms. While the database engine is able to internally optimize JOIN queries, we cannot profit from these optimization mechanisms due to the pseudonymized data scheme where the links between the tables are literally broken up, i.e., all critical relations are hidden by encryption. Using SSDs instead of traditional disk-based hard drives considerably improved overall performance.

Due to the calculation performances of HSMs, cryptography only added a minor overhead to the overall retrieval, as did the actual pseudonymization algorithm, which was rather modest in terms of computational requirements. At the client side, the smart card performance as cryptographic environment proved to be a limiting factor. While the authentication operation is executed only once for each session, session key encryption had a considerable impact on the overall retrieval operation, especially when encrypting documents. Thus, host-based encryption, i.e., workstation instead of smart card, is more appropriate (either SK encryption/decryption at the host or TLS). As the critical inner symmetric key is still kept secure within the HSM, the impact in security is therefore minimal.

6. Threat scenarios

We identified multiple threat scenarios that need to be addressed by security mechanisms of the PIPE framework. Security is typically verified by evaluation against the security attributes confidentiality (C), integrity (I), and availability

data model and the workflows can be easily adapted to a true 1:1 relationship between identification and health record to support more complex CDA documents with unique header sections.

Table 2 – Threat scenarios.

Attack scenario	Security attribute	Level of protection
Cryptanalysis	C, I	Full
Traffic interception/manipulation	C, I, A	Full
Man-in-the-middle/impersonation	C, I, A	Full
Social engineering	C	Partial
Hardware theft	C, A	Full
Database theft	C	Full
Unauthorized data retrieval	C	Full
Malicious code	C, I	Partial

(A) [51]. As privacy is paramount in PIPE, we primarily focus on confidentiality, and, in second place, on integrity. The scenarios are summarized in Table 2.[5]

As we rely on standard cryptographic algorithms that can be easily replaced when required, we assume perfect cryptography. Authentication with two random nonces and session key encryption protects against traffic interception and man-in-the-middle attacks in untrusted communication environments. Because the smart cards are protected by the user-owned PIN and the secret keys currently in use are automatically erased from the HSM when removed from its power source, hardware theft is not an issue. Database theft cannot be fully prevented but information leakage is limited by the pseudonymized storage structure. Unauthorized data retrieval, i.e., knowledge of the correct link between identification and health record, is prevented by encryption with the users' inner symmetric keys. Social engineering cannot be technically prevented, but its consequences are limited by the user-based encryption. Malicious code is a serious issue, especially when the pseudonymization logic is compromised. The tamper resistance of the smart card and the HSM provides a reasonable level of security and ensures that the secret keys cannot be leaked, but the pseudonymization server and the workstation still need to be protected against malicious code with malware scanners with up-to-date signatures.

7. Conclusion

Electronic health records allow the structured and expandable collection of medical data to be shared between authorized health care providers to improve the quality of patients' treatment and to reduce costs. The problem with electronically stored health data are the arising privacy concerns, especially with centrally stored health data. The disclosure of sensitive patient information may result in serious problems for the patient, including harassment or denied services. In this paper, we have presented a novel protocol for the centralized pseudonymization of health data and provided an overview of the concept and a detailed description of the protocol. This approach allows the unlinked storage of medical data and related patient-identifying information, while still making it possible to reestablish the patient/health data relationship for authorized users. We carried out a detailed security verification comprising scenario-based threat analysis and prototyping to show that the methodology provides a high level of security through the application of a dedicated hardware security module for the secured cryptographic operations.

[5] Details on the threat scenarios can be found as supplementary material on the publisher's website.

Summary points

What is known:

- The availability of reliable information has a tremendous impact on decisions regarding the patients' care.
- Electronic health records allow the improvement of communication between health care providers and of their access to data and documentation, leading to better clinical and service quality.
- The electronic storage of health data raises considerable privacy concerns.

What this work adds:

- Identifies shortcomings of existing approaches and methodologies.
- Provides a methodology for the pseudonymization of medical data to guarantee patients' privacy.
- The proposed PIPE architecture allows primary and secondary use of the data at the same time.

Author contributions

The manuscript was written jointly by the both authors, being a result of a 2,5 years research project.

Conflict of interest statement

There are NO conflicts of interest.

Acknowledgments

This work was supported by grants of the Austrian Government's BRIDGE Research Initiative (contract 824884) and was performed at the research center Secure Business Austria funded by the Federal Ministry of Economy, Family and Youth of the Republic of Austria and by the City of Vienna.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijmedinf.2010.10.016.

REFERENCES

[1] S. Märkle, K. Köchy, R. Tschirley, H.U. Lemke, The PREPaRe system—patient oriented access to the personal electronic medical record, in: Proceedings of the 17th International Congress and Exhibition on Computer Assisted Radiology

and Surgery, ser. International Congress Series, no. 1256, 2001, pp. 849–854.
[2] F.R. Ernst, A.J. Grizzle, Drug-related morbidity and mortality: updating the cost-of-illness model, Journal of the American Pharmacists Association 41 (2) (2001) 192–199.
[3] United States Department of Health Human Service, HIPAA administrative simplification: enforcement; final rule, Federal Register/Rules and Regulations 71 (32) (2006).
[4] 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, Official Journal of the European Communities L 281 (1995) 31–50.
[5] Council of Europe, European Convention on Human Rights, Martinus Nijhoff Publishers, 1987.
[6] S. Fischer-Hübner, IT-Security and Privacy: Design and Use of Privacy-Enhancing Security Mechanisms, Springer, 2001.
[7] S. Hinde, Privacy legislation: a comparison of the US and European approaches, Computers and Security 22 (5) (2003) 378–387.
[8] G. Hornung, C.F.-J. Goetz, A.J.W. Goldschmidt, Die künftige Telematik-Rahmenarchitektur im Gesundheitswesen, Wirtschaftsinformatik 47 (2005) 171–179.
[9] U.S. Department of Health & Human Services Office for Civil Rights, "Summary of the HIPAA Privacy Rule", 2003. [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/.
[10] U.S. Congress, "Health Insurance Portability and Accountability Act of 1996", 104th Congress, 1996. [Online]. Available: http://www.cms.hhs.gov/HIPAAGenInfo/Downloads/HIPAALaw.pdf.
[11] T. Schabetsberger, E. Ammenwerth, G. Göbel, G. Lechleitner, R. Penz, R. Vogl, F. Wozak, What are functional requirements of future shared electronic health records? Connecting Medical Informatics and Bio-Informatics (2005) 1070–1075.
[12] B. Riedl, T. Neubauer, G. Goluch, O. Boehm, G. Reinauer, A. Krumboeck, A secure architecture for the pseudonymization of medical data, in: Proceedings of the Second International Conference on Availability Reliability and Security, 2007, pp. 318–324.
[13] R.C. Barrows, P.D. Clayton, Privacy, confidentiality, and electronic medical records, Journal of the American Medical Informatics Association 13 (1996) 139–148.
[14] J Attridgde, An Overview of Hardware Security Modules, SANS Institute, Tech. Rep., 2002.
[15] J. Montagnat, F. Bellet, H. Benoit-Cattin, V. Breton, L. Brunie, H. Duque, Y. Legré, I.E. Magnin, L. Maigne, S. Miguet, J.M. Pierson, L. Seitz, T. Tweed, Medical images simulation, storage, and processing on the European DataGrid Testbed, Journal of Grid Computing 2 (4) (2004) 387–400.
[16] R Sharp, Principles of Protocol Design, Springer, 2008.
[17] A. Pfitzmann, M. Koehntopp, Anonymity, unobservability and pseudeonymity—a proposal for terminology, in: International Workshop on Designing Privacy Enhancing Technologies, Springer-Verlag, Inc., New York, 2001, pp. 1–9.
[18] K.A. Taipale, Technology, security and privacy: the fear of Frankenstein, the mythology of privacy and the lessons of King Ludd, International Journal of Communications Law & Policy 9 (2004).
[19] A Rector, J. Rogers, A. Taweel, D. Ingram, D. Kalra, J. Milan, P. Singleton, R. Gaizauskas, M. Hepple, D. Scott, R. Power, Clef—joining up healthcare with clinical and post-genomic research, in: Proceedings of UK e-Science All Hands Meeting, 2003, pp. 203–211.
[20] C. Thielscher, M. Gottfried, S. Umbreit, F. Boegner, J. Haack, N. Schroeders, Patent: data processing system for patient data, Int. Patent, WO 03/034294 A2, 2005.
[21] K. Maris, The human factor in information technology, in: Proceedings of Hack.lu, Luxembourg, 2005.
[22] T. Thornburgh, Social engineering: The "Dark Art", in: Proceedings of the First Annual ACM Conference on Information Security Curriculum Development, ACM Press, 2004, pp. 133–135.
[23] M. Bishop, D. Gollmann, J. Hunker, C.W. Probst, Countering insider threats, in: Dagstuhl Seminar Proceedings 08302, 2008.
[24] K. Pommerening, Medical requirements for data protection, in: Proceedings of IFIP Congress, vol. 2, 1994, pp. 533–540.
[25] K. Pommerening, M. Reng, Secondary use of the electronic health record via pseudonymisation Medical and Care Compunetics, vol. 1, IOS Press, 2004, pp. 441–446.
[26] A.A.E. Kalam, Y. Deswarte, G. Trouessin, E. Cordonnier, A generic approach for healthcare data anonymization, 2004.
[27] R. Noumeir, A. Lemay, J. Lina, Pseudonymization of radiology data for research purposes, 2007.
[28] Digital Imaging Communications in Medicine, National Electrical Manufacturers Association Std., 2008.
[29] R.L. Peterson, Patent: encryption system for allowing immediate universal access to medical records while maintaining complete patient control over privacy, US Patent US 2003/0074564 A1, 2003.
[30] J. Caumanns, Der Patient bleibt Herr seiner Daten: Realisierung des eGK-Berechtigungskonzepts über ein ticketbasiertes, virtuelles Dateisystem, Informatik-Spektrum 29 (5) (2006) 323–331.
[31] Fraunhofer Institut, Spezifikation der Lösungsarchitektur zur Umsetzung der Anwendungen der elektronischen Gesundheitskarte, March 2005.
[32] C.D. Stingl, D. Slamanig, D. Rauner-Reithmayer, H. Fischer, Realisierung eines sicheren zentralen Datenrepositories, in: Tagungsband, DACH Security, 2006, pp. 1–15.
[33] C. Stingl, D. Slamanig, Berechtigungskonzept für ein e-health-portal, in: G. Schreier, D. Hayn, E. Ammenwerth (Eds.), eHealth 2007—Medical Informatics Meets eHealth, no. 227, Österreichische Computer Gesellschaft, 2007, pp. 135–140.
[34] C. Stingl, D. Slamanig, Privacy aspects of e-health, in: Proceedings of the Third International Conference on Availability, Reliability and Security, 2008, pp. 1226–1233.
[35] M.K. Bond, Understanding Security APIs, Ph.D. dissertation, University of Cambridge, Computer Laboratory, Emmanuel College, 2004.
[36] Federal information processing standards publication, "Security requirements for cryptographic modules (Fips pub 140-2)," Institute of Standards and Technology (NIST), Tech. Rep., 05 2001.
[37] R. Anderson, M. Bond, J. Clulow, S. Skorobogatov, Cryptographic Processors—A Survey, University of Cambridge, Computer Laboratory Tech. Rep., 2005.
[38] PKCS#11 v2.20: Cryptographic Token Interface Standard, RSA Laboratories Std., 2004.
[39] D.C. Wherry, Secure Your Public Key Infrastructure with Hardware Security Modules, SANS Institute, Tech. Rep., 2003.
[40] M. Lorch, J. Basney, D. Kafura, A hardware-secured credential repository for grid PKIs, in: Proceedings of the 4th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2004.
[41] T. Rössler, H. Leithold, R. Posch, E-voting: a scalable approach using XML and hardware security modules, in: Proceedings of the 2005 IEEE International Conference on e-Technology, e-commerce and e-Service EFF'05, 2005.
[42] A. Baldwin, S. Shiu, Hardware encapsulation of security services, in: 8th European Symposium on Research in

Computer Security (ESORICS), ser. LNCS, vol. 2808, Springer, Berlin/Heidelberg, 2003.
[43] M. Casassa-Mont, A. Baldwin, J. Pato, Secure Hardware-based Distributed Authorisation Underpinning a Web Service Framework, Trusted Systems Laboratory, HP Laboratories Bristol, Tech. Rep., 2003.
[44] A. Baldwin, S. Shiu, Hardware security appliances for trust, in: First International Conference on First International Conference, ser. LNCS, vol. 2692, Springer, 2003.
[45] A. Ferreira, S. Shiu, A. Baldwin, Towards accountability for electronic patient records, in: Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems (CBMS'03), 2003, pp. 189–194.
[46] A. Shamir, How to share a secret, Communications of the ACM 22 (11) (1979) 612–613.
[47] T. Jurgensen, S. Guthery, Smart Cards: The Developer's Toolkit, Pearson Education, Inc. Prentice Hall PTR, Upper Saddle River, 2002.
[48] B. Holcombe, Government Smart Card Handbook, U.S. General Services Administration (GSA), 2004.
[49] International Statistical Classification of Diseases and Related Health Problems (ICD), World Health Organization Std., 2007.
[50] R.H. Dolin, L. Alschuler, C. Beebe, The HL7 clinical document architecture, Journal of the American Medical Informatics Association 8 (6) (2001) 552–569.
[51] A. Avizienis, J.-C. Laprie, B. Randell, C. Landwehr, Basic concepts and taxonomy of dependable and secure computing, IEEE Transactions on Dependable and Secure Computing 1 (1) (2004) 11–33.