A Legal Evaluation of Pseudonymization Approaches

Thomas Neubauer
Vienna University of Technology
Vienna, Austria
neubauer@ifs.tuwien.ac.at

Mathias Kolb
Secure Business Austria
Vienna, Austria
kolb@securityresearch.ac.at

***Abstract*—Privacy is one of the fundamental issues in health care today and a fundamental right of every individual. Several laws were enacted that demand the protection of patients' privacy. However, approaches for protecting privacy often do not comply with legal requirements or basic security requirements. This paper highlights research directions currently pursued for privacy protection in e-health and evaluates common pseudonymization approaches against legal criteria taken from Directive 95/46/EC and HIPAA. Thereby, it supports decision makers in deciding on privacy systems and researchers in identifying the gaps of current approaches for privacy protection as a basis for further research.**

***Keywords*-security, privacy, pseudonymization, e-health**

I. Introduction

Privacy is a trade-off between the patient's demands for privacy as well as the society's need for improving efficiency and reducing costs of the health care system. Electronic health records (EHR) improve communication between health care providers and access to data and documentation, leading to better clinical and service quality [2]. The EHR promises massive savings by digitizing diagnostic tests and images (cf. [3]). The pervasiveness of electronic devices has resulted in the almost constant surveillance of everyone and the permanent storage of personal data that is used and analyzed by corporations or intelligence services. With informative and interconnected systems comes highly sensitive and personal information that is often available over the Internet and – what is more concerning – hardly protected. It is a fundamental right of every individual to demand privacy because the disclosure of sensitive data may cause serious problems for the individual. Insurance companies or employers could use personally identifiable information to deny health coverage or employment. Although a variety of laws were enacted that demand the protection of privacy, only a few of the existing approaches comply with the current legal requirements. The individuals' rights are difficult and costly to pursue because they are limited in the absence of a dedicated authority to oversee and enforce compliance. The disclosure of personal data may be avoided through the use of privacy enhancing technologies (PET), such as anonymization, or more importantly, pseudonymization. Whereas anonymity allows unlinkability and maybe unobservability, it prevents any useful two-way communication. Pseudonymization ensures that a user may use a resource without disclosing his identity, but can still be accountable for that use [4].

This paper presents an evaluation of six current privacy enhancing technologies that specifically aim at protecting medical data by using pseudonymization and, thus, are used as a basis for EHR systems. The paper answers two major questions: (i) Which pseudonymization approaches adhere to the current privacy laws and (ii) what are the major drawbacks of current pseudonymization approaches. In the scope of this paper we regard evaluation as the "systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy" (cf. [5]). Based on the categorization of House [6] and Stufflebeam & Webster [7] we use a combination of objectivist approaches: The Testing programs approach and the Objectives-based approach. The objectives used for the evaluation are taken from the legal acts HIPAA and the EU Directive. This evaluation provides management decision makers such as chief privacy officers and chief security officers with a funded decision-making basis for the selection of privacy-enhancing technologies in the e-health area. As literature does not provide evaluations focusing on the comparison of PETs in e-health in literature so far, this paper provides a major contribution to the research area of privacy.

II. Legal Background

Nowadays, society is collecting all kinds of information. In daily life, several types of information are tracked, which are highly sensitive and can even be damaging to individuals and organizations [8][9][10]. For example, the supermarket tracks which items have been bought, mobile phone providers keep track of customer movements, airlines know what type of seat and meal is preferred and hotel chains keep records of room preferences. The exchange and storage of this information became very cheap and simple over the Internet. For this reason it is more important than ever to protect the privacy of individuals. In more than 30 countries, privacy laws protect the data of individuals [11]. The content of these privacy laws varies in each country, but they are mostly based on the Organization for Economic

Cooperation and Development (OECD) Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [12].

Throughout history, collected information of individuals has been abused in several ways. Regarding the individual's privacy, historically the phrase "to be let alone", defined at the US Supreme Court in 1834, became famous. In the years during World War II, the German government abused census data to identify people of certain ethnic, religious or other targeted groups (cf. [13][14]. As various states gained in power and size, the first privacy laws were introduced in order to protect minorities. In 1948 the United Nations ratified a right to privacy in article 12 of the Universal Declaration of Human Rights. The UN declaration defines privacy as "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation". UN member countries are morally, if not legally, bound by such declarations. Everyone has the right to the protection of the law against such interference or attacks. A citizen's right of privacy is also recognized in the Article 8 of the European Convention for the Protection of Human Rights and Fundamental Freedoms from 1950. In 1966 a Computer Bill of Rights was suggested, followed by a Rights to Privacy Act in 1967 was proposed, which banned wiretapping and electronic eavesdropping.

The first national data-protection law was passed 1973 in Sweden, followed by the United States in 1974 and West Germany in 1977 [15]. In the United States, privacy has not gained much political attention. Discussions on privacy have been driven often by events in Europe. In the 1970s, concerns over privacy reached new heights, because of the abuse of wiretapping, tax, bank and telephone records during the Watergate scandal [13]. These concerns gave birth to the Privacy Act of 1974, which applies only to records of personal information held by federal agencies. These agencies are allowed to keep records only if relevant and necessary. They are not allowed to create secret files of an individual without giving the right to copy their own files. Furthermore, agencies are not permitted to disclose these records without the agreement of the individual - except within the agency for routine use or law enforcement [13].

By the end of the seventies more and more European States had passed privacy laws. To spread these laws across Europe, the Organization for Economic Cooperation and Development (OECD) published the Guidelines on the Protection of Privacy and Transborder Flows of Personal Data and the Convention of the Council of Europe in 1980/1981 that defined the provisions for the protection of individuals with regard to the automatic processing of personal data. To protect private electronic communications from unauthorized access by the government, the Electronic Communications Privacy Act of 1986 and the Computer Matching and Privacy Protection Act of 1988 have been introduced in the US. There are currently no privacy acts in the US that could be compared to the European acts. There are a handful of laws which cover the use of private data in health care [17][18][19], the electronic commerce industry [20], the cable-television industry [21] and a few other areas. A definition of personal data is given in Section 8(8) of the Online Privacy Protection Act (OPPA) [22]:

> '... information collected online from an individual that identifies that individual, including first and last name, home and other physical address, e-mail address, social security number, telephone number, any other identifier that the Commission determines identifies an individual, or information that is maintained with, or can be searched or retrieved by means of, data described above ...'

In 1995 the European Union (EU) passed the Data Protection Directive (95/46/EC) [23]. This directive applies to all personal data, which is collected or processed either electronically or in old-fashioned paper-filing systems. Article 2(a) of the Data Protection Directive (95/46/EC) defines personal data as:

> '... any information relating to an identified or identifiable natural person (data subject); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity ...'

Moreover, the Data Protection Directive (95/46/EC) is based on eight principles to which all data controllers are subject. These principles limit the usage of collected personal data [24][23][25]:

1) The data must be processed fairly and lawfully.
2) The data must be collected for explicit and legitimate purposes and used accordingly.
3) The data must be accurate and where necessary, kept up to date.
4) Organizations have to provide mechanisms to correct, delete or block data.
5) The data that identifies individuals must not be kept longer than necessary.
6) The data must be processed in accordance with the rights of the data subject.
7) Every organization must ensure the security and integrity of personal data, that they are processing.
8) It is not permitted to transfer personal data outside the European Union unless the country ensures an adequate level of protection

Furthermore, to ensure fair and lawful processing of the collected data, the data controller has to inform the data subjects which data will be collected and used. The individual also must be informed of the type of third parties the collected data will be disclosed to and the data subject

must have the option to decline [23][25][26][24]. Especially sensitive data like in the health care sector need more privacy protection than non-sensitive data in other sectors. Sensitive medical data like the state of medical health, for example being HIV positive or having chronic illness, could harm a person if they are accessed by unauthorized persons. For example, an employer who accesses medical data of her employees, could use this information to dismiss an employee. Another example could be an insurance company denying a contract because of a chronic illness.

In the European Union, the Data Protection Directive (95/46/EC) [23] already implements protection for sensitive data, which are related to racial and ethnic background, political affiliation, religious or philosophical beliefs, trade-union membership, sexual preferences and health [23][25][13][24]. Besides this Data Protection Directive, an additional Working Document [27] has been released by the Article 29 Working Party of the European Union, which provides guidelines for the interpretation of the data protection legal framework for EHR systems and explains some of the general principles. The Working Document also gives indications on the data protection requirements for setting up EHR systems, as well as for the applicable safeguards. The processing of sensitive data is generally prohibited but is tolerated under specific circumstances [25]. Some of these circumstances are:

- if the data subject explicitly agrees on the processing of her sensitive data.
- if the processing of data is allowed by law.
- if the subject is unable to agree on the processing, e.g., due to unconsciousness.

Furthermore the Protection Directive (95/46/EC) defines the rights for the individual. Some of these rights are:

- to receive information about the processing of their own data,
- to receive a copy of all personal data held by the data controller,
- the prevention of direct marketing and automated decision-making,
- to seek damages for breach of the data protection principles.

In 2006 the United States Department of Health and Human Service Health issued the Health Insurance Portability and Accountability Act (HIPAA) which demands the protection of patients data that is shared from its original source of collection (cf. [16][17][18][19][28][29][30]). It is based on five principles:

1) Consumer control of medical information,
2) Boundaries that limit disclosure of medical treatment and
3) Payment accountability for violation of patient's rights with specific federal penalties,
4) Public responsibility for protecting public health, conducting medical research, improving quality of care and fighting health care fraud or abuse, and
5) Security of health information by organizations entrusted with that information.

The five principles only apply to individually identifiable health information, which is:

- created by or received from health care providers, employers or the clearinghouse.
- related to the provision of health care or the past, present or future medical condition.
- identifies or could reasonably be used to identify an individual.
- has been transmitted electronically or maintained in any other form or medium.

However, the act does not include other medical data, for example car insurance that has medical coverage or general sickness absence in the workplace that is not the subject of the health plan [24].

The disclosure of Protected Health Information (PHI) is permitted in certain cases. For example, the data is disclosed to the individual itself, the data is de-identified to carry out health plan's own treatment, payment or health care operations. Furthermore, the data owner could give consent to the processing of her medical data. To protect the privacy of individuals, many rights have been set up under the Health Insurance Portability and Accountability Act. Individuals have the right:

- to inspect or copy their own information,
- to request amendment or correction of erroneous or incomplete information,
- to request the restriction of use or disclosure,
- to give authorization for certain uses and disclosures.

III. Description of Pseudonymization Approaches

This chapter describes current pseudonymization approaches in detail. Thereby, we differentiate between three approaches. Firstly, there is the plain-text approach in which all data is readable for everyone. This approach could be compared with the traditional paper-record system. Secondly, there is the encrypted-text approach in which all data are encrypted and only accessible to persons with the key to decrypt this data. Thirdly, there is the pseudonymization approach, in which only the reference between the data and the data owner is encrypted. Table I gives an overview of the approaches:

A. Peterson Approach

Peterson [32] claims to provide a system for providing personal medical information records to an individual without jeopardizing privacy. The main ideas behind the approach are (i) the encryption of patient's data, (ii) the universal access to medical records by any (also unauthorized)

Description	*Name*	*References*
Plain-text approach	Approaches of Pommerening	[31]
Encrypted-text	Approach of Peterson	[32]
	Elektronische Gesundheits Karte	[33][35][36][34][37][38][39][40]
Pseudonymization approach	Pseudonymization of Information for	
Privacy in e-Health	[41][42][43][44][45]	
	Approach of Thielscher	[46]
	Approach of Slamanig and Stingl	[47][48][49]

Table I
OVERVIEW OF PSEUDONYMIZATION APPROACHES

person while (iii) the patient is responsible for granting privacy.

The user registers at the provider's website, receives a unique Global Key (GK) and server side key ($SSID$) generated by the provider and has to provide a unique Personal Encryption Key (PEK) as well as a password. The server returns a unique global key GK, which has to be different from the PEK. GK, PEK and password are stored in the Data Table. The user is demanded to enter a PEK until he provides a unique one. After registration the user may print the GK on an ID Card (on paper).

This approach consists of three database tables, the user table, the security table and the personal data table. The user table contains the GK, the PEK, a password and a foreign key to the security table. The security table contains a primary key, the method of encryption for the PEK, a server side encryption key and method and a foreign key to the personal medical data table. This table contains a primary key and the data, which is double encrypted with the PEK and the server side encryption key. Data is stored double encrypted in the database. If the user wants to retrieve data from the database, the user enters the GK or PEK, which are sent to the server through a firewall and checked if they match any entry in the database. The user enters an arbitrarily key and gets immediate access to the records without authentication.

In case of an emergency, the health care personnel can retrieve the medical data of the patient by entering the global key GK, or if the patient is responsive to verbal commands, she can tell them the private encryption key PEK. The system looks up the database for the entered GK or PEK and returns the decrypted medical data. The server looks up the $SSID$ and all corresponding data table row numbers needed for retrieving the (medical) data entries from the database. The records are decrypted using (i) the PEK and the personal encryption method and (ii) the server side encryption key $SSEK$ and the server side encryption method and delivered to the user. To modify or delete this medical data, the patient has to enter her password, which has been provided at registration time.

Table II shows the different access levels of this approach. If a person knows the global key GK or PEK or both, but does not have a password, she is able to view medical data sets. To be able to add, modify or delete medical datasets, the person has to provide the password. Peterson argues, that these access levels protect patient's privacy, because the data does not contain any identifying information. So, for an attacker, it would be of no interest to receive anonymous data.

Global Key	*Personal Key*	*Password*	*Resulting Action*
No	No	No	Access Denied
Yes	No	No	View Only
No	Yes	No	View Only
Yes	Yes	No	View Only
No	No	Yes	Access Denied
Yes	No	Yes	View and Edit
No	Yes	Yes	View and Edit
Yes	Yes	Yes	View and Edit

Table II
APPROACH OF PETERSON: ACCESS LEVELS [32]

B. Pseudonymization of Information for Privacy in e-Health (PIPE)

PIPE (cf. [50][42][51][52]) is a architecture that provides the following contributions compared to other methodologies: PIPE allows (i) the authorization of health care providers or relatives to access defined medical data on encryption level, (ii) provides a secure fall-back mechanism, in case the security token is lost or worn out, (iii) stores the data without the possibility of data profiling, and (iv) provides secondary use without establishing a link between the data and its owner.

The client is a service, which provides an interface to legacy applications, manages requests to local smart card readers and creates a secure connection to the server. The server, also called Logic (L), handles requests from clients to the storage. The data in the storage is divided into two parts, the personal data and the pseudonymized medical data. The link between personal data and pseudonymized medical data is protected through a hull-architecture. The hull-architecture (see Figure 1) contains a minimum of three

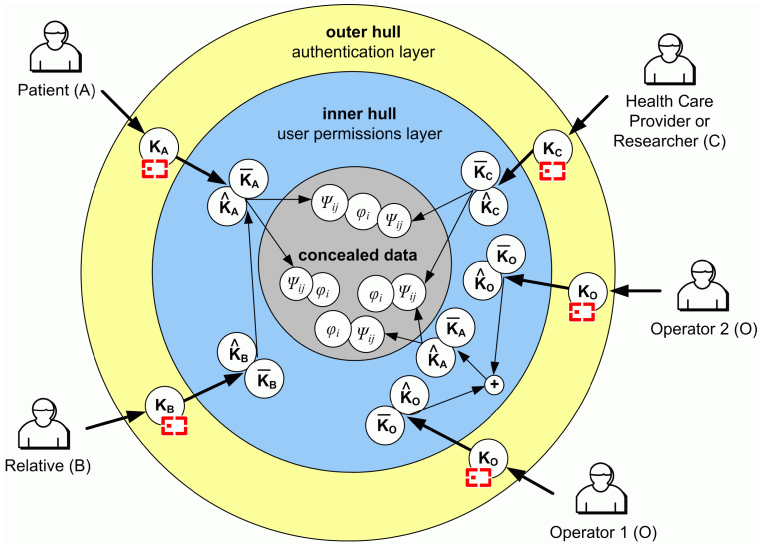

Figure 1. PIPE: Layered model representing the authorization mechanism

security-layers: the authentication layer (outer hull), the user permission layer (inner hull) and the concealed data layer. To reach the next hull, there are one or more secrets, for example, symmetric or asymmetric keys or hidden relations, in every hull-layer. A definition of all system attributes can be found in table III. PIPE defines users with different roles comprising patient A, relative B, health care provider C and operator O. The patient is the owner of her data and has full control of her datasets. She is able to view her medical data, add and revoke health care providers and she may define relatives, who have the same rights as herself. Health care providers can be authorized by the patient to see and create subsets of anamnesis data. The operators provide a backup in case the token needs to be replaced.

- The authentication layer contains an asymmetric key pair, e.g., the patient's outer public key K_A and outer private key K_A^{-1}. These keys are stored on a smart card and are protected with a pin code. The outer private key is used to decrypt the keys of the permission hull-layer.
- The permission layer contains an asymmetric key pair and a symmetric key, e.g., the patient's inner public key $\widehat{K}_A$, inner private key $\widehat{K}_A^{-1}$ and symmetric key $\overline{K}_A$. The symmetric key is encrypted with the inner private key and is used to en-/decrypt pseudonyms in the concealed data layer. If a patient associates a relative, her inner private key $\widehat{K}_A^{-1}$ is encrypted with the relative's inner public key $\widehat{K}_B$. So, the relative is able to decrypt the patient's symmetric key $\overline{K}_A$ with her inner private key $\widehat{K}_B^{-1}$, until the patient's inner private key $\widehat{K}_A^{-1}$ is changed.
- The concealed data layer contains hidden relations, which are called pseudonyms. Each medical data set is associated with one or more pseudonyms ψ_{i_j}. As the patient is the owner of her medical data and the person with security clearance, she owns the so called root-pseudonym ψ_{i_0}. These pseudonyms are calculated with an algorithm, which is based on a secret key. In our case, this secret key is the symmetric key of the user. Only instances, who are able to decrypt one of these pseudonyms ψ_{i_j}, can rebuild the link between the patient and her medical data.

To find the pseudonyms to rebuild the link to the medical data, the authors introduced keywords. Keywords are selected on creation time of the medical data or when another user is authorized. They are encrypted with the symmetric key of the root user and the user, who is being authorized. After the keywords are stored in the database, the user can select any of this keywords to find the pseudonym.

C. Electronic health card (eGK) architecture

The electronic health card architecture [33][34][35][36][37][38][39][40] is an approach of the Fraunhofer Institute supported by the Federal Ministry of Health Germany. The EGK is designed as a service-oriented architecture (SOA) with some restrictions: The health card can only be accessed locally on the client side. Services should use remote procedure calls for communication due to performance and availability issues. Therefore, the system architecture is divided into five layers:

- The *presentation* layer defines interfaces to communicate with the user,
- the *business logic* layer combines different services, which are processed automatically,
- the *service* layer provides special functional uncoupled services,
- the *application* layer realizes the user right and data management, and

	Patient	*Relative*	*HCP*	*Operator*	*Logic*
abbreviation	A	B	C	O	L
unique identifier	A_{id}	B_{id}	C_{id}	O_{id}	
(outer public key, private key)	(K_A, K_A^{-1})	(K_B, K_B^{-1})	(K_C, K_C^{-1})	(K_O, K_O^{-1})	(K_L, K_L^{-1})
(inner public key, private key)	$(\widehat{K}_A, \widehat{K}_A^{-1})$	$(\widehat{K}_B, \widehat{K}_B^{-1})$	$(\widehat{K}_C, \widehat{K}_C^{-1})$	$(\widehat{K}_O, \widehat{K}_O^{-1})$	
inner symmetric key	$\overline{K}_A$	$\overline{K}_B$	$\overline{K}_C$	$\overline{K}_O$	$\overline{K}_L$
key share				$\sigma_\iota(K)$	
medical data / anamnesis	φ_i				
pseudonym	ψ_{i_j}				

Table III
PIPE: DEFINITION OF SYSTEM ATTRIBUTES

- the *infrastructure* layer contains all physical hardware and software management, for example, data storage, system management, virtual private networks, etc.

With this layered architecture, the system provides several service applications such as emergency data, electronic prescription, electronic medical report or a electronic health record system. The system includes a ticketing concept to realize some uncoupled action in combination with security mechanisms, to comply with the privacy policy: All data, which will be stored in the virtual file system is encrypted with a one-time symmetric key, called session key. This session key is encrypted with the public key of the patient. To decrypt the data, the patient has to decrypt the session key with his private key and finally the data will be decrypted with this session key. A user is authenticated by using a Challenge-Response approach. Therefore the system generates a random number. This number will be encrypted with the public key of the user. Only the user is allowed to decrypt this random number with the private key, which is stored on her health card and can send it back to the eGK system. Furthermore, the ticketing concept manages the access rights to the system. A file or directory in this virtual file system has a default ticket-toolkit and any amount of private ticket-toolkits, called t-node (see Figure 2). The user defines a private ticket-toolkit for every other user in the system. This private ticket-toolkit could have stronger or looser access policies as the default ticket-toolkit. The ticket-toolkit contains a ticket-building tool, a ticket-verifier, the access policy list and a encrypted link to the directory or file. Every user holds a root directory in the virtual file system, which does not have a parent node. Furthermore, any directory contains unencrypted links to the ticket-toolkits of their child nodes. This technique enables the system to perform a fast selection of sub nodes (select * from t-nodes where parentID = directoryID).

To be able to find the root node of a specific user, the query service maps a unique identifier, for example the insurance number to the internal user and returns a ticket-toolkit containing a encrypted link to the root node. If there is no private ticket-toolkit available for the user, who performed the request, the system returns a default ticket-toolkit, which is based on a challenge. If the user is able to solve this challenge, she will get the access rights, which have been defined in the default access policy. Both, the hybrid encryption and the challenge response technique are based on the asymmetric key pair, which is stored on the patients' health card. Neither the operating company nor any public administration organization could recover the data, which has been stored in the system, if the patient lost the smart card or the card is worn out. To overcome this problem, the eGK architecture optionally provides the possibility to store a second private ticket-toolkit for every entry. This private ticket-toolkit uses an asymmetric key pair, which is stored on an emergency card. The architecture does not specify this emergency card, but recommends to use the card of a family member or a notary.

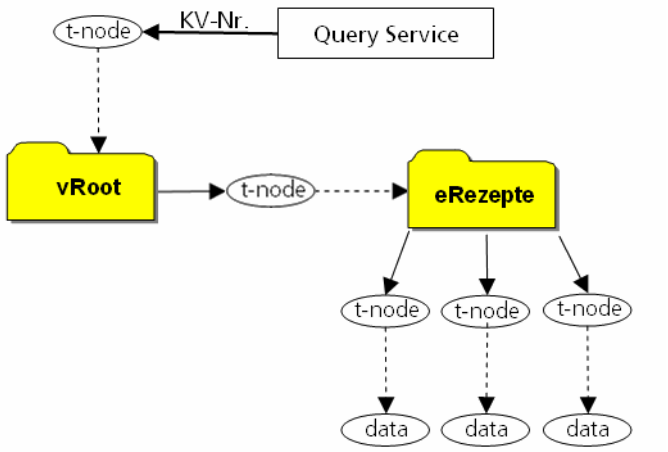

Figure 2. eGK: Virtual file system [33]

D. Thielscher Approach

Thielscher [46] proposes an electronic health record system, which uses decentralized keys stored on smart cards. The medical data are split into identification data and the anamnesis data and stored into two different databases. The key stored on the smart card of a patient is used to link the patient identity to her datasets. Therefore, this key generates a unique data identification code (DIC), which is also stored

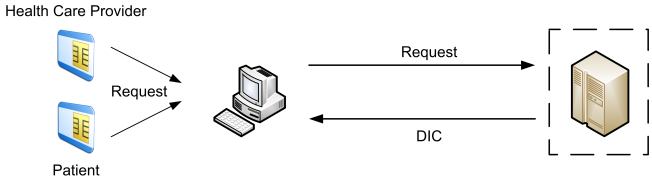

Figure 3. Thielscher: Architecture [46]

in the database. Such a DIC does not contain any information to identify an individual. Data identification codes are shared between the patient and health care providers to authorize them to access the medical data set. For more security the authorization is limited to a certain time period. After this period any access attempt is invalid. The system provides a mechanism in case of an emergency. Some parts of the patient's individual health data is stored directly on the smart card. A health professional has immediate access to this data in case of an emergency. Moreover, the system includes an emergency call center which is authorized to access the central database for requests and to read the data in case of an emergency. Therefore, the health professional has to confirm their identity to the call center.

E. Approach of Slamanig and Stingl

Stingl and Slamanig [48][49] propose a concept for an e-health portal with a public user repository and a document repository with encrypted medical documents. The link between these repositories is realized by a 5-tuple authorization concept $(U_S, U_R, U_C, U_P, D_i)$, which contains the identifiers of the sender U_S, the receiver U_R, the data creator U_C, the concerning user (i.e. patient) U_P, and the document reference with the decryption key D_i. All tuples except for the receiver are encrypted with the receiver's public key. Authorizations and the amount of disclosed information depend on the tuples used:

- $(U_1, U_1, U_1, U_2, D_1)$: User 1 creates this tuple concerning user 2 for accessing document 1.
- $(U_1, U_3, -, -, D_1$: User 1 authorizes user 3 to access document 1 without disclosing information on the data creator and the concerning user.
- $(U_1, U_2, U_1, U_2, D_1)$: User 1 authorizes the concerning user 2 to access document 1 disclosing himself as the data creator.

In order to provide unlinkability, each user has a set of sub-identities, realized as independently chosen pseudonyms, with individual asymmetric keypairs. One of these sub-identities is defined as public identity used for authorizations, while the others are kept secret. Upon receipt of an authorization, the recipient first decrypts the relation with the private key of the public sub-identity, replaces the receiver tuple with one of his secret sub-identities, and then reencrypts the remaining tuples with the corresponding

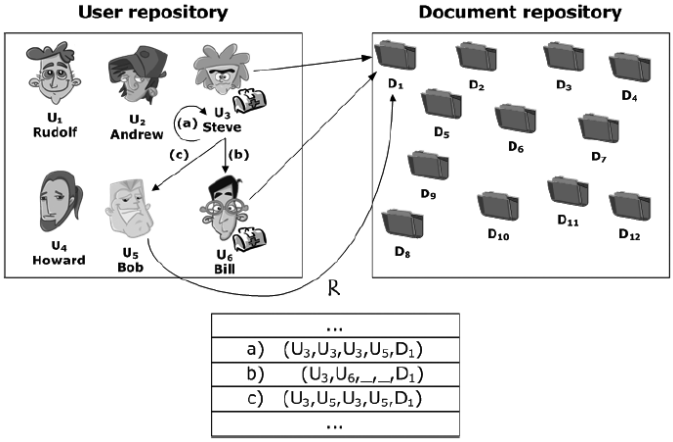

Figure 4. Slamanig and Stingl: Repositories and Shares [49]

public key such that the authorization tuple cannot be identified by any observer, except for the corresponding user. To prevent so called *disclosure attacks* (users forced to disclose their medical data, e.g., at job interviews) a special sub-identity can be chosen which includes only non-critical data. Highly sensitive data can be hidden in another sub-identity [54]. As fall-back mechanism, the authors mentioned that the distributed key backup to N users using a (t, N)-threshold secret sharing scheme could be implemented, because the users private keys are essential for the system.

The authors also propose the application of techniques such as anonymous authentication and obfuscation to further improve the patients' privacy. Obfuscation can be realized by intentionally producing collisions when selecting pseudonyms such that the pseudonyms are not unique, obfuscating the exact links between pseudonyms and documents. But obfuscation produces computational overhead because of invalid returned tuples (tuples actually not possessed by the user need to be identified as such by decrypting them and checking their semantic content). Anonymous authentication provides unlinkability between individual access operations but needs to be executed for each transaction individually.

In [55] and [56] they propose the application of their concept for personal health records (PHR). The medical documents are organized in virtual folders (where the content does not need not be disjunct) which in turn are controlled by sub-identities. In addition to identity pseudonymization, the folders and documents are pseudonymized as by foreign key encryption such that the documents, folders, and sub-identities cannot be linked by an observer [48].

Anonymous authentication and pseudonymization in the form of sub-identities provide a great deal of unobservability, both from the static and dynamic viewpoint. The usage of a special sub-identity managing only non-critical information also prevents exposure of sensitive data as a result of disclosure attacks. While reencryption of the authorization

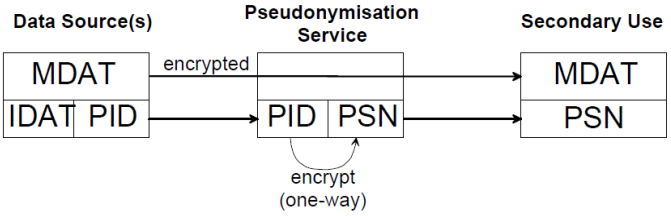

Figure 5. Pommerening: Data Flow for One-Time Secondary Use [31]

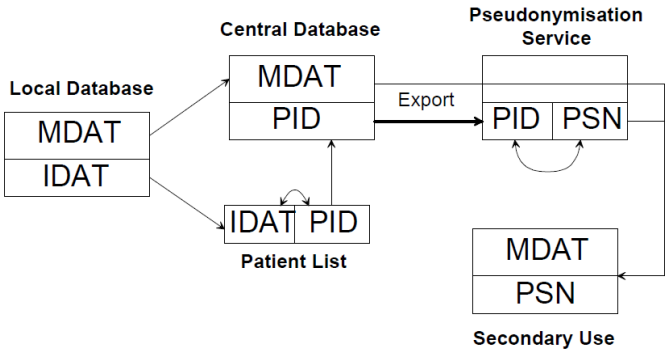

Figure 6. Pommerening: Data Flow for many Secondary Uses [31]

tuple after receipt ensures unobservability, it also prevents that the sender can revoke this authorization. In fact, the sender cannot control if the recipient authorizes a third person without the consent of the data owner. For finding a particular document, the user needs to select the correct sub-identity and folder and decrypt all document references (and document information) to determine the desired document; a query mechanism is not provided. Finally, because of the fully encrypted documents, secondary use is not possible without decryption by an authorized user.

F. Pommerening Approaches

Pommerening [31] proposes different approaches for secondary use of medical data. He differs between one-way and reversible pseudonyms. The first approach is based on data from overlapping sources for one-time secondary use. In this case, overlapping sources could be, e.g., data from different EHRs or biomaterial banks, which have been collected on another examination. To connect the data, a unique identifier (PID) is introduced. Figure 5 shows the pseudonymization workflow. A pseudonymization service encrypts the PID with a hash algorithm, and the medical data (MDAT) is encrypted with the public key of the secondary user. The secondary user can decrypt the medical data and merge the data of a person, but cannot identify it.

The second approach is also based on one-time secondary use, but with the possibility to re-identify the patient. Therefore, Pommerening extends the first approach with a PID service, which stores a reference list containing the identity of the patient (IDAT) and the associated PIDs. In case the patient should be notified, the pseudonymization service decrypts the pseudonym (PSN) and sends the request to the PID service, which notifies the data source owner.

The third approach fits the need of a research network with numerous secondary users. It supports long-term observation, e.g., of a patient with chronic diseases and allows to send research results to the patient or her responsible health care provider. The export and pseudonymization procedure is shown in figure 6. Therefore a physician exports her local database to the central researcher database. The identification data will be replaced with a PID in the PID service. For each secondary use the data will be exported through the pseudonymization service. The PID is encrypted by the pseudonymization service with a project specific key to ensure that different projects get different pseudonyms.

IV. Legal Evaluation

Pseudonymization approaches (e.g., used for securing electronic health record systems) have to adhere certain requirements to accord with privacy laws in the European Union or United States. The following set of requirements has been extracted from the Directive 95/46/EC of the European Parliament (DPA) and the Health Insurance Portability and Accountability Act (HIPAA) (cf. [23][24][53][17][18][19]).

- *User authentication*: The system has to provide adequate mechanisms for user authentication. This could be done, for example with smart cards or finger print.
- *Data ownership*: The owner of the medical data has to be the patient. The patient should be able to define who is authorized to access and create her medical records.
- *Limited access*: The system must ensure that medical data is only provided to authenticated and authorized persons.
- *Protection against unauthorized and authorized access*: The medical records of an individual have to be protected against unauthorized access. This includes system administrators who should not be able to access these medical records, for example, through compromising the database.
- *Notice about use of patients data*: The patient should be informed about any access to her medical records.
- *Access and copy own data*: The system has to provide mechanisms to access and copy the patients own data.
- *Unobservability*: Pseudonymized medical data should not be observable and linkable to a specific individual in the system.
- *Secondary use*: The system should provide a mechanism to export pseudonymized data for secondary use and a possibility to notify the owner of the exported data, if new medicaments or treatment methods are available.

Legal Requirements	*DPA*	*HIPAA*	*PIPE*	*eGK*	*Po*	*Pe*	*Th*	*St*
User authentication	x	x	x	x	-	o	x	x
Data ownership	x	x	x	x	-	-	x	o
Limited access	x	x	x	x	o	-	x	x
Protection against unauthorized and authorized access	x	x	x	x	o	-	o	x
Notice about use of patients data	x	x	x	x	-	-	-	-
Access and copy own data	x	x	x	x	o	x	x	x
Unobservability	x	x	x	x	x	-	x	x
Secondary use	-	x	x	o	x	-	-	-

Table IV
EVALUATION OF PSEUDONYMIZATION APPROACHES

Abbreviations		
Po	...	approach of Pommerening
Pe	...	approach of Peterson
Th	...	approach of Thielscher
Sl	...	approach of Slamanig and Stingl

Legend for DPA and HIPAA		
x	...	defined and accurate with the law
-	...	undefined in the law

Legend for pseudonymization approaches		
x	...	fully implemented
o	...	partially implemented
-	...	not implemented

Table V
LEGEND FOR TABLE IV

Table IV applies the legal criteria defined above to the selected pseudonymization approaches. Characteristics that are accurate with the law or fully implemented are denoted with x, whereas characteristics that are not accurate with the law or not implemented are denoted with $-$ and o indicates properties that are partially implemented.

Fulfilling legal requirements is an important precondition in order to guarantee security. However, since legal requirements are often defined in a generic way they leave room for interpretation. This results in a variety of approaches that are often vulnerable to typical attack scenarios. Table VI presents a list of typical attack scenarios and evaluates these criteria against the pseudonymization approaches described earlier.

- *Insider abuse*: Medical personnel may abuse their access rights for their own purposes. For example, they may want to know how family members or celebrities are being treated [57]. Insiders do not only abuse their privileges for their own purposes, they may release information to outsiders for spite, revenge or profit [57].
- *Social engineering*: is a common method to get information about a person. Therefore, an attacker could bribe or mislead an administrator of the pseudonymization system. For example, the attacker could fake her identity to get a new security token.
- *Data Disclosure*: Data mining attacks are a major threat for the disclosure of sensitive data as shown by Sweeney (cf. [58]). Sweeney was able to combine medical data with an electronic version of a city's voter list. The attacker can collect statistics and information about the data. In the worst case scenario the attacker could reconstruct the pseudonyms.
- *Attacker deletes data*: If an attacker breaks into the system, she may have the possibility to delete data. Therefore the system should be able to detect such changes and inform the system administrator about this attack and request a restoration of the datasets.
- *Attacker modifies data*: An attacker, who has broken into the system, may also change some datasets. Therefore, the system should digitally sign all records in order to detect modifications.
- *Attacker authorizes internal users*: An attacker could try to authorize an internal user or herself to be able to gain access to medical data of other users.
- *Attacker authorizes external users*: An attacker could try to authorize an external user or herself to be able to gain access to medical data of other users.
- *Administrator accesses data*: Administrators of the

Possible security issues	*PIPE*	*eGK*	*Po*	*Pe*	*Th*	*Sl*
Insider abuse	-	-	x	x	x	-
Social engineering	o	o	x	x	x	o
Data Disclosure	-	-	o	x	o	-
Attacker deletes data	o	x	x	x	x	o
Attacker modifies data	-	-	x	x	x	o
Attacker authorizes internal users	-	-	o	x	-	x
Attacker authorizes external users	-	-	o	x	-	-
Administrator accesses data	-	-	x	x	-	-
Administrator accesses cryptographic keys	-	-	o	x	o	-

Table VI
COMPARISON OF SECURITY ISSUES AND PSEUDONYMIZATION APPROACHES

Abbreviations		
Po	...	approach of Pommerening
Pe	...	approach of Peterson
Th	...	approach of Thielscher
Sl	...	approach of Slamanig and Stingl

Legend for pseudonymization approaches		
x	...	security issue
o	...	possible security issue
-	...	no security issue

Table VII
LEGEND FOR TABLE VI

pseudonymization system could access the database if the data is pseudonymized only by disclosure.

- *Administrator accesses cryptographic keys*: If system administrators have access to the private keys of individuals, she may have the possibility to decrypt all pseudonyms and link anamnesis to individuals. Every attacker who gets administration privileges could steal the database containing the keys.

Most of the approaches implement the requirements of *user authentication*, *data ownership*, *limited access* and serve control mechanisms *against unauthorized and authorized access*. The implementation of the requirement *protection against unauthorized and authorized access* is inadequate. Additional requirements, which enhance the security of the system and the containing datasets, are widely implemented. The approaches of Pommerening and Peterson only pseudonymize data on export. The approaches of Pommerening have the drawback that the generated pseudonyms from the PID service are stored in a reference patient list, to be able to re-build the link to the patient. To enhance the security, this list can be stored at a third party institution, but this measure does not prevent an abuse of the list through an insider of the third party institution. The system permits attackers to steal the database with all data linked to individuals. Moreover, system administrators could abuse their access privileges to release information to outsiders for revenge, profit or their own purposes [57]. An attacker could bribe an insider of the third party institution to get access to the patient list or the identifying data of some pseudonyms. The Peterson approach has some major security issues. Although the data is doubly encrypted an attacker getting access to the database gets access to all data stored on the server because the keys needed for decrypting the data are (i) also stored in the same database and (ii) what is even more important the relation between the tables (thus between the identification data and the medical data) are stored in clear text. An attacker getting access to the database can decrypt all information and, as the password is stored in the database as well as the keys, the attacker may change data stored in the database. The PEK is selected by the user but must be unique in the system. This behavior does not only open a security leak because the user trying to chose a key is informed about the keys that already existing in the system. An attacker could use the keys reported as existing for immediate access to the medical data associated with this key. Moreover, this behavior is impractical and inefficient in practice as the user might have to select dozens of keys before he enters a valid one. Peterson tried to prevent the following types of attacks: insider abuse, disclosure of weakly pseudonymized data and databases being stolen. He did so by defining that no identifiable data is allowed to be stored. However, the system is not able to check if identifiable words exist in the

data. Thielscher's approach comes with the shortcoming, that the pseudonyms are stored centrally in the patient mapping list for recovery purposes. To prevent attacks to this list, Thielscher keeps this list off-line, but this mechanism cannot prevent insider abuse or social engineering attacks. The usage of a patients-pseudonyms list as fall-back mechanism could lead to security issues. The work-around of Thielscher to keep the patients-pseudonyms list off-line promises a higher level of security, but does not prevent the system against social-engineering or insider attacks. Furthermore, it does not provide protection if the attacker gets physical access to the computer. Another drawback of the system is the emergency call center. This call center can abuse their access privileges to get access to medical data of any patient. The drawback of the approach of Slamanig and Stingl is that an attacker (a person who gets access privileges on a document) may authorize other users, send faked medical documents or disclose medical data. For example, the requirements to send a faked medical document are, (i) access to the database, (ii) the public pseudonym U_P of the user, which the attacker wants to harm, (iii) any public pseudonym to fake the sender U_S and creator U_C, (iv) the public pseudonym and the public key K_R of the receiver U_R, for example the employer, and (v) a harmful document D_i. After the attacker has all the required information, she inserts a new tuple into the authorization table. After the next login of the receiver, the system replaces the public pseudonym of the user with a private pseudonym of the receiver. The authors suggest obfuscation to handle this problem. The approach does not prevent tuple reordering and, thus, allows the attacker to modify data.

PIPE, eGK and Slamanig/Stingl store the data pseudonymized in the database. Attackers who get access to the database or system administrators cannot link the data to individuals. All those approaches provide a high level of security. Even if the attacker breaks into the database, she would not be able to link and read the stored data. Maybe, the attacker could do a data profiling attack and get some informations from the unencrypted keywords, if these contain any identifiable words. The only way to link the data to an individual is by doing a social engineering attack and fake the identity of the person, the attacker wants to attack. Therefore, the attacker would have to fake a official photo identification in order to get a new smart card to access the system. Another method to link data to an individual is by doing a data mining or data profiling attack.

V. Conclusion

Health care require the sharing of patient related data in order to provide efficient patients' treatment. As highly sensitive and personal information is stored and shared within highly interconnected systems (e.g., electronic health records), there is increasing political, legal and social pressure to guarantee patients' privacy. Although, legislation demands the protection of patients' privacy, most approaches that lay claim to protect patients' privacy fail in fulfilling legal requirements.

This paper gave an overview of research directions that are currently pursued for privacy protection in e-health and evaluated common pseudonymization approaches against legal criteria taken from legal acts and literature. Thereby, this paper answered the questions (i) which pseudonymization approaches adhere to the current privacy laws and (ii) what are the major drawbacks of pseudonymization approaches. In order to answer the first research question, seven legal requirements have been extracted from relevant legal acts. These requirements could be used for the future development of pseudonymization approaches. At the moment, only two out of the six evaluated pseudonymization approaches fulfill the legal requirements. Therefore, only two out of the six approaches can actually be considered for use in the European Union and United States. Moreover, the results of the evaluation show that newer approaches already consider legal demands and fulfill more legal requirements of the European Union and the United States. An additional security evaluation, carried out to answer the second research question, shows that there are major drawbacks in most of the systems. Some approaches use a pseudonym-patient mapping list, which could very easily be abused by an insider of the system, for example a system administrator. A more secure way was presented by eGK, where all data is linked to backup security tokens. However, if both security tokens are accidentally destroyed, for example by fire, all data would be lost forever. Only two approaches suggest a solution to share the keys of the security token in the system using a threshold scheme. PIPE is the only approach which implements such a fall-back mechanism.

From the six candidates that were evaluated, only two can be seriously considered for use in practice. The result show that more contemporary approaches fulfill more of the legal requirements of the European Union and the United States. Whereas the eGK approach encrypts patients' data, PIPE leaves the decision of encrypting patients' data up to the user. Therefore, PIPE turns out to be the more appropriate option if secondary use is demanded. Apart from this difference both approaches - eGk and PIPE - provide a similar level of security and fulfill the majority of the applied criteria. The results of the evaluation can support decision makers (such as chief security officers) especially in health care in their decision process when it comes to the selection of a system for protecting patients' data according to legal requirements posed by HIPAA or the EU Directives. Furthermore, the results may assist researchers in identifying the gaps of current approaches for privacy protection as a basis for further research.

VI. Acknowledgments

This work was supported by grants of the Austrian Government's FIT-IT Research Initiative on Trust in IT Systems under the contract 816158 and was performed at Secure Business Austria, a competence center that is funded by the Federal Ministry of Economy, Family and Youth of the Republic of Austria as well as by the provincial government of Vienna.

References

[1] Thomas Neubauer and Mathias Kolb. Technologies for the pseudonymization of medical data: A legal evaluation. In *Proceedings of the IEEE International Conference on Systems (ICONS)*, 2009.

[2] S. Märkle, K. Köchy, R. Tschirley, and 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*, number 1256 in International Congress Series, pages 849–854, 2001.

[3] Frank R. Ernst and Amy J. Grizzle. Drug-related morbidity and mortality: Updating the cost-of-illness model. Technical report, University of Arizona, 2001.

[4] Common criteria for information technology security evaluation, ISO/IEC 15408:1999.

[5] C. H. Weiss. *Evaluation: Methods for studying programs and policies*. Prentice Hall, 2nd edition, 1998.

[6] E. R. House. Assumptions underlying evaluation models. *Educational Researcher*, 7(3):4–12, 1978.

[7] D. L. Stufflebeam and W. J. Webster. An analysis of alternative approaches to evaluation. *Educational Evaluation and Policy Analysis*, 2(3):5–19, 1980.

[8] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, 1998.

[9] Bruce Schneier. Risks of data reuse. Schneier on Security - Blog, June 2007. Last access 28.09.2009.

[10] Bruce Schneier. Our data, ourselves. Schneier on Security - Blog, May 2008. Last access 28.09.2009.

[11] Alfred Kobsa. Personalized hypermedia and international privacy. *Commun. ACM*, 45(5):64–67, 2002.

[12] Organisation for Economic Cooperation and Development (OECD). Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data c(80)58/final, 1980.

[13] Solveig Singleton. Privacy and Human Rights: Comparing the United States to Europe. In *The Future of Financial Privacy*, pages 186–201, 1999.

[14] William Seltzer. Population Statistics, the Holocaust, and the Nuremberg Trials. *Population and Development Review*, 24(3):511–552, 1998.

[15] Colin John Bennett. *Regulating Privacy: Data Protection and Public Policy in Europe and the United States*. Cornell University Press, 1992. ISBN: 0801480108.

[16] United States Department of Health & Human Service. HIPAA Administrative Simplification: Enforcement; Final Rule. *Federal Register / Rules and Regulations*, 71(32), 2006.

[17] U.S. Department of Health & Human Services Office for Civil Rights. Summary of the HIPAA Privacy Rule, 2003.

[18] U.S. Department of Health & Human Services Office for Civil Rights. Your Health Information Privacy Rights.

[19] U.S. Congress. Health Insurance Portability and Accountability Act of 1996. *104th Congress*, 1996.

[20] Federal Trade Commission. Children's online privacy protection act. United States federal law, 15 U.S.C. §6501-6506, October 1998.

[21] U.S. House of Representatives. U.S. Code - Title 47 - Telegraphs, Telephones, and Radiotelegraphs - Chapter 5 - § 551.

[22] H. R. 84. Online privacy protection act of 2005. 109th Congress, 1st Session, Bill, October 2005. This bill never became law.

[23] 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 281:31–50, 1995.

[24] Stephen Hinde. Privacy legislation: A comparison of the US and European approaches. *Computers and Security*, 22(5):378–387, 2003.

[25] *Data protection in the European Union - Citizen Guide*. European Union, 2001.

[26] Gerhard Steinke. Data privacy approaches from US and EU perspectives. *Telematics and Informatics*, 19(2):193–200, 2002.

[27] European Union, Article 29 Working Party. Working document on the processing of personal data relating to health in electronic health records, February 2007.

[28] Tim Churches. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Medical Research Methodology*, 3(1), 2003.

[29] George J. Annas. Hipaa regulations - a new era of medical-record privacy? *The new england journal of medicine*, 348(15):1488–1490, 2003.

[30] David Baumer, Julia Brande Earp, and Fay Cobb Payton. Privacy of medical records: IT implications of HIPAA. *ACM SIGCAS Computers and Society*, 30(4):40–47, 2000.

[31] Klaus Pommerening and Michael Reng. *Medical And Care Compunetics 1*, chapter Secondary use of the Electronic Health Record via pseudonymisation, pages 441–446. IOS Press, 2004.

[32] Robert 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.

[33] Fraunhofer Institut. Spezifikation der Lösungsarchitektur zur Umsetzung der Anwendungen der elektronischen Gesundheitskarte, March 2005.

[34] Jörg Caumanns, Herbert Weber, Arne Fellien, Holger Kurrek, Oliver Böhm, Jan Neuhaus, Jörg Kunsmann, and Bruno Struif. Die eGK-Lösungsarchitektur Architektur zur Unterstützung der Anwendungen der elektronischen Gesundheitskarte. *Informatik-Spektrum*, 29(5):341–348, 2006.

[35] Jörg Caumanns. Der Patient bleibt Herr seiner Daten: Realisierung des eGK-Berechtigungskonzepts über ein ticketbasiertes, virtuelles Dateisystem. *Informatik-Spektrum*, 29(5):323–331, 2006.

[36] Andreas Rottmann. CAMS dirigiert die eGK. *Datenschutz und Datensicherheit - DuD*, 30:153–154, 2006.

[37] Jan Neuhaus, Wolfgang Deiters, and Markus Wiedel. Mehrwertdienste im Umfeld der elektronischen Gesundheitskarte. *Informatik-Spektrum*, 29(5):332–340, 2006.

[38] Bernd Blobel and Peter Pharow. Wege zur elektronischen Patientenakte. *Datenschutz und Datensicherheit - DuD*, 30(3):164–169, 2006.

[39] Gerd Bauer. Aktive Patiententerminals. *Datenschutz und Datensicherheit - DuD*, 30(3):138–141, 2006.

[40] D. Wilhelm, A. Schneider, and C. F. J. Götz. Die neue Gesundheitskarte. *Der Onkologe*, 11(11):1157–1165, 2005.

[41] Bernhard Riedl, Thomas Neubauer, and Oswald Boehm. Patent: Datenverarbeitungssystem zur Verarbeitung von Objektdaten. *Austrian-Provisional-Application, Application No. A 1928/2006*, 2006.

[42] Bernhard Riedl, Thomas Neubauer, Gernot Goluch, Oswald Boehm, Gert Reinauer, and Alexander Krumboeck. A secure architecture for the pseudonymization of medical data. In *Proceedings of the Second International Conference on Availability, Reliability and Security*, pages 318–324, 2007.

[43] Bernhard Riedl, Veronika Grascher, and Thomas Neubauer. Applying a threshold scheme to the pseudonymization of health data. In *Proceedings of the 13th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC'07)*, pages 397–400, 2007.

[44] Bernhard Riedl, Veronika Grascher, Stefan Fenz, and Thomas Neubauer. Pseudonymization for improving the privacy in e-health applications. In *Proceedings of the Forty-First Hawai'i International Conference on System Sciences*, page 255, 2008.

[45] Bernhard Riedl, Veronika Grascher, Mathias Kolb, and Thomas Neubauer. Economic and Security Aspects of the Appliance of a Threshold Scheme in e-Health. In *Proceedings of the Third International Conference on Availability, Reliability and Security*, pages 39–46, 2008.

[46] Christian Thielscher, Martin Gottfried, Simon Umbreit, Frank Boegner, Jochen Haack, and Nikolai Schroeders. Patent: Data processing system for patient data. *Int. Patent, WO 03/034294 A2*, 2005.

[47] Christian Stingl, Daniel Slamanig, Dominik Rauner-Reithmayer, and Harald Fischer. Realisierung eines sicheren zentralen Datenrepositories. In *Tagungsband DACH Security*, 2006.

[48] Christian Stingl and Daniel Slamanig. Berechtigungskonzept für ein e-Health-Portal. In Günter Schreier, Dieter Hayn, and Elske Ammenwerth, editors, *eHealth 2007 - Medical Informatics meets eHealth*, number 227, pages 135–140. Oesterreichische Computer Gesellschaft, 2007.

[49] Daniel Slamanig and Christian Stingl. Privacy aspects of ehealth. In *Proceedings of the Third International Conference on Availability, Reliability and Security*, pages 1226–1233, 2008.

[50] Bernhard Riedl, Thomas Neubauer, and Oswald Boehm. Patent: Datenverarbeitungssystem zur Verarbeitung von Objektdaten. *Austrian Patent, Nr. 503291, September*, 2007.

[51] Bernhard Riedl, Veronika Grascher, and Thomas Neubauer. A secure e-health architecture based on the appliance of pseudonymization. *Journal of Software*, 3:23–32, 2008.

[52] Thomas Neubauer and Bernhard Riedl. Improving patients privacy with pseudonymization. In *Proceedings of the International Congress of the European Federation for Medical Informatics*, number 136 in Studies in Health Technology and Informatic, pages 691–696, 2008.

[53] Gerrit Hornung, Christoph F.-J. Götz, and Andreas J. W. Goldschmidt. Die künftige Telematik-Rahmenarchitektur im Gesundheitswesen. *Wirtschaftsinformatik*, 47:171–179, 2005.

[54] Daniel Slamanig and Christian Stingl. How to preserve patient's privacy and anonymity in web-based electronic health records. In *Proceedings of the 2nd Internatinal Conference on Health Informatics (HEALTHINF 2009)*, 2009.

[55] Daniel Slamanig and Christian Stingl. Sophisticated methods to prevent insider attacks against PHR systems. In *Proceedings of the IADIS International Conference on e-Health*, 2009.

[56] Daniel Slamanig and Christian Stingl. Ein sicheres patientenzentriertes konzept für personal health records. In *eHealth 2009 - Health Informatics meets eHealth*, 2009.

[57] Thomas C. Rindfleisch. Privacy, information technology, and health care. *Commun. ACM*, 40(8):92–100, 1997.

[58] Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.