Privacy-Preserving Storage and Access of Medical Data through Pseudonymization and Encryption

Johannes Heurix[1] and Thomas Neubauer[2]

[1] SBA Research, Austria
jheurix@sba-research.org
[2] Vienna University of Technology
Institute of Software Technology and Interactive Systems, Austria
neubauer@ifs.tuwien.ac.at

Abstract. E-health allows better communication between health care providers and higher availability of medical data. However, the downside of interconnected systems is the increased probability of unauthorized access to highly sensitive records that could result in serious discrimination against the patient. This article provides an overview of actual privacy threats and presents a pseudonymization approach that preserves the patient's privacy and data confidentiality. It allows (direct care) primary use of medical records by authorized health care providers and privacy-preserving (non-direct care) secondary use by researchers. The solution also addresses the identifying nature of genetic data by extending the basic pseudonymization approach with queryable encryption.

Keywords: e-Health, Privacy, Pseudonymization

1 E-Health and the Need for Privacy

Today's health care is driven by the goal of streamlining and optimizing processes in order to reduce costs without compromising the quality of patient treatment. E-health denotes the application of information and communication technologies (ICT) to support the medical workflows and to improve the communication between health care providers. Over the past years, interconnected systems, such as electronic health records (EHR), provide the technical infrastructure for facilitated document sharing by making them digitally available, having the potential to increase the quality of health care while keeping the costs at a controlled level [1]. However, facilitated access also means higher chance of misuse. Thus sensitive information such as HIV infection data or drug abuse histories must be adequately protected to prevent discrimination, such as denied insurance coverage. Even the sole probability of developing a serious illness may be sufficient to decide against health or live insurance coverage. A particular example of this form of prejudice is called genetic discrimination, the biased treatment of people due to gene mutations that may cause or increase the risk of an inherited disorder [4], [2]. There are numerous documented cases where the results of so-called

predictive genetic tests were disclosed to insurance companies resulting in denied insurance coverage, although genetic tests usually deliver uncertain probabilities instead of clear-cut predictions of developing a genetic disorder. Genetic discrimination is also an issue with job applications and employment, where employees were fired because of 'unfavorable' genetic tests and thus keeping them would be too 'risky'.

Although legal acts such as the Genetic Information Nondiscrimination Act (GINA) [3], the Health Insurance Portability and Accountability Act (HIPAA) [12], and the Directive 95/46/EC [5] by the EU exist, technical solutions are still required to prevent the disclosure of medical records to unauthorized persons. At the same time, the vast amounts of digitized data produced in today's health care environment should be available for secondary use, for non-direct care use of personal health information including (but not limited to) analysis and research, as well as quality and safety measurement [9]. Providing access to this rich source of information can help to expand knowledge about diseases and treatment and enhance the effectiveness and efficiency of health care, which in turn improves direct care for the individual patient. But considering reports on buying and selling of non-anonymized patient and health care provider data by the medical industry without the explicit consent from patients or physicians, making these data available poses a significant privacy risk. The effective primary and secondary use of medical records is a major challenge for developing appropriate privacy protection measures.

1.1 Anonymization and Encryption

Two techniques often mentioned when confidentiality and privacy of data is required are *anonymization* and *encryption*. Anonymization refers to removing the identifier from the medical data such that the records cannot be traced back to the corresponding patient [11]. Anonymization can be achieved by depersonalization, the removal of any patient-identifying information from the health records. Because perfect depersonalization, where the data subject is no longer identifiable at all circumstances, is practically impossible to achieve, the assumption can be relaxed to modifying the health data such that the corresponding patient can either not at all or only with a 'disproportionate amount of time, expense and labour' be identified (cf. [6]). A well-known technique of anonymization is k-anonymity [10] where identifying information is removed in such a way that each person cannot be distinguished from at least k-1 individuals by comparing the remaining data stored in the database. A particular downside of anonymization is the fact that it cannot be reversed, which means that anonymized health data cannot be used for direct care or primary use where the link between health data and corresponding patient obviously needs to be known by the health care providers. Anonymization also has its downsides in secondary use, where it is usually applied: As the patient cannot be identified any more, they cannot be contacted to ask for necessary further information or be directly informed of any results either, thus cannot immediately profit from advances in medical treatment. Anonymization may also be inadequate for securely storing genetic data

due to their identifying nature. The other technique, data encryption, is usually employed when data confidentiality is required. By fully encrypting health data with a secret key only known to the patient, his or her privacy can be assured as well. Native data encryption is provided by many major database providers and prevents unauthorized disclosure of any sensitive data as long as the decryption key is kept secret and protected adequately. Unlike anonymization, full data encryption is obviously reversible, but the major problem is that secondary use of the records in research projects is entirely prevented, unless the patient explicitly decrypts the data, thus unconcealing his or her identity. Also considering the technical heterogeneous environment of health institutions, (authorized) sharing of encrypted records is also more complicated. Furthermore, encryption and decryption can be very time-consuming when large (monolithic) medical records are involved such as imaging data, in this case rendering data access operations quite tedious.

1.2 Pseudonymization as a Solution

Pseudonymization combines the strengths of anonymization and (full) document encryption: It achieves unlinkability by introducing specifiers (pseudonyms) which cannot be associated with the patient without knowing a certain secret. Other than plain anonymization, it is reversible. Therefore, with prior depersonalization of health records, it allows storing the records in an anonymized state, while this anonymity can be reversed by authorized persons having the knowledge of the secret key. While pseudonymization itself also relies on cryptography (when no cleartext mapping/linking list is involved), only metadata need to be encrypted, and thus the necessary cryptographic overhead can be considerably reduced, compared to simply fully encrypting the health documents.

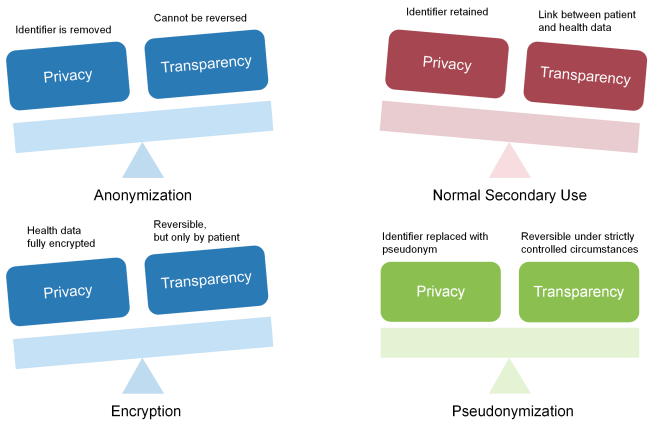

Fig. 1. Trade-off between Privacy and Transparency

Figure 1 represents the difficulty of keeping the patient's privacy and data usability as a trade-off between privacy and transparency: Both anonymization and encryption shift the emphasis on privacy, compromising transparency, while secondary use without anonymization or data encryption discloses the link between patient and health data, compromising patient's privacy. Pseudonymization however is able to keep the balance between privacy and transparency.

2 The PIPE Approach

We developed an approach denominated as 'Pseudonymization of Information for Privacy in E-Health' (PIPE) to provide a secured and privacy-preserving storage and retrieval of sensitive medical data [7]. The basic idea is that many medical records alone such as X-ray images (of frequent mishaps like broken legs) are insufficient to uniquely identify the patient after depersonalization. Therefore, this medical content is separated from identifying information (patient's name, address, ...) and both records are assigned randomly-selected pseudonyms, i.e., identification and health pseudonyms that form a 1:1 relation. These pseudonyms act as access tokens: knowing the correct pseudonyms allows to relink health records to the corresponding patients. The pseudonyms are protected by encryption with a user-specific secret key. We developed the approach with the following aspects in mind:

- Privacy-Preserving Storage: Unlike some other pseudonymization approaches where data are not pseudonymized until export, health data are already stored uncoupled from the identification record, i.e., the data are protected even against potential internal attackers having/gaining direct access to the database (e.g., administrator).
- Privacy-Preserving Secondary Use: The decoupled storage structure facilitates privacy-preserving secondary use, e.g., for research institutions without taking additional anonymization steps. Still, the capability for relinking the records for authorized users allows direct care primary use.
- Patient-Centric Authorization Model: In our approach, the patient is defined as data owner who retains full control over his health data at all times, i.e., the patient is the only person that is able to define access authorizations for trusted persons. In this context, *authorizations* and *access* do not refer to traditional access rights but to the ability to relink certain medical records to the patient and can be stated for just specific health records (e.g., for health care providers) or for the entirety of the patient's data (e.g., for close relatives).
- Secured Authentication: Because passwords are often too weak for dependable authentication, we utilize security tokens (i.e., micro controller smartcards with integrated crypto chips) for dependable authentication. The token acts as trusted user-owned cryptographic module.
- Cryptographic Standards: For simple implementation, we use standard cryptographic symmetric and asymmetric protocols (e.g., RSA and AES) which can be easily replaced should the need arise.

2.1 User Roles

The major roles in our pseudonymization scheme are the patients, health care providers, and trusted relatives:

- Patient: As data owner, the patient is the only person that is in full control over his or her health data and can add and delete owned data at his or her discretion. The patient can also grant data access authorizations for trusted health care providers and relatives. The patient retrieves his or her records by root pseudonyms which are created for and (initially[3]) only known to the patient.
- Health Care Provider: Authorizations for a health care provider involve access rights for specifically selected health records defined by the patient. The patient creates an authorization by creating and assigning a unique (shared) pseudonym to the particular health record to share with the health care provider. Combined with another new pseudonym assigned to the patient's identification record, the newly created pseudonyms form the access token which can be deleted when necessary. The pseudonym pair is known to both patient and health care provider, i.e., encrypted with both their keys. A health care provider can also be authorized to add a new health record for the patient. In this case, the health care provider is automatically granted access to the record in the future as well.
- Relative: In contrast to health care providers, a relative is granted access to the entirety of a patient's data records by sharing the secret information to decrypt the root pseudonyms. The relative is also automatically authorized to access any records stored in the future as well, either added by the patient himself/herself or by a trusted and authorized health care provider.

The pseudonymized data structure of PIPE provides two different 'views' depending on the granted authorizations (cf. Figure 2). The left side represents the data view for administrators and secondary users (unauthorized in terms of data confidentiality and privacy), or internal/external malicious users. Although the identification and health records are clearly visible for them, they are not able to identify the correct links between the pseudonymized health records and identification records. All they can do is try to guess. Authorized users however, i.e., the patient, authorized health care providers, and relatives, are able to 'see through' the pseudonymization and can re-establish the correct links. As shown on the right side, it becomes clear that the four highlighted medical records belong to the patient represented by the identification record in the middle.

2.2 Security Architecture

We built PIPE around a security architecture with three different layers each responsible for different aspects of the security framework (cf. Figure 3): The *authentication layer* is realized by the outer asymmetric keypair (outer private

[3] cf. Relative

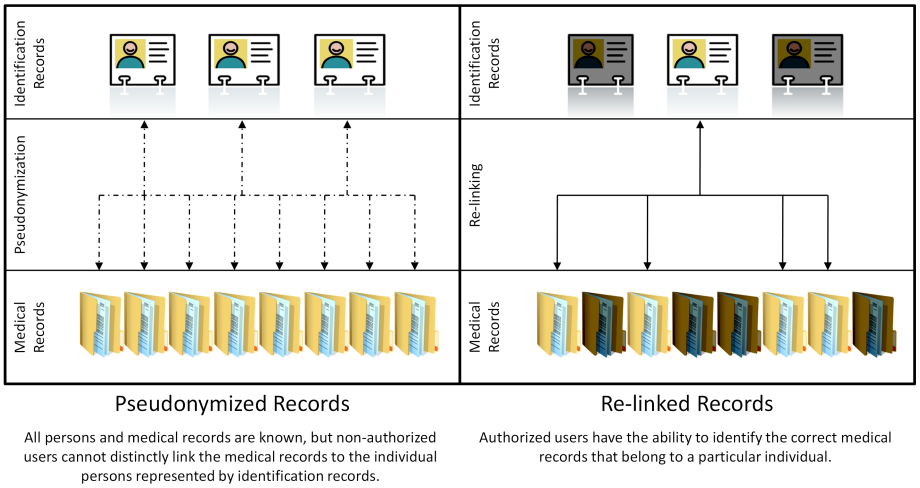

Fig. 2. Pseudonymized View for Authorized and Unauthorized Persons

and public keys). The purpose of the first layer is to unambiguously identify the user, as the outer private key is only stored at the user's security token. With this outer private key, the next layer can be accessed by decrypting the inner private key which in turn allows decryption of the inner symmetric key. The inner asymmetric keypair and the inner symmetric key form the *authorization layer*. The inner symmetric key grants the user access to the final and innermost layer, the *pseudonymized data layer*. By decrypting the pseudonyms with the inner symmetric key, the user can finally relink the health record to the corresponding identification record.

In Figure 3, the health care provider on the right is authorized for a specific health record which is related to an identification record (dotted arrow) representing the patient in the center. This relation is mirrored by the relation between the identification and health pseudonyms, protected by encryption with the inner symmetric keys. Only users in the possession of the correct security tokens are able to pass each layer to finally decrypt this relation. The shared pseudonyms are encrypted with both the patient's and the health care provider's inner symmetric keys. The root pseudonyms are solely encrypted with the patient's inner symmetric key and therefore initially known to the patient only. Because of sharing the patient's inner private key with the relative who stores his version of this key encrypted with his inner symmetric key, the relative thus gains access to the root pseudonyms too.

2.3 Search within Pseudonymized Records

Because the encrypted pseudonyms do not (and must not) contain any semantic information on the data records they are referenced with, a query mechanism is required. We developed a simple keyword mechanism for that purpose. As

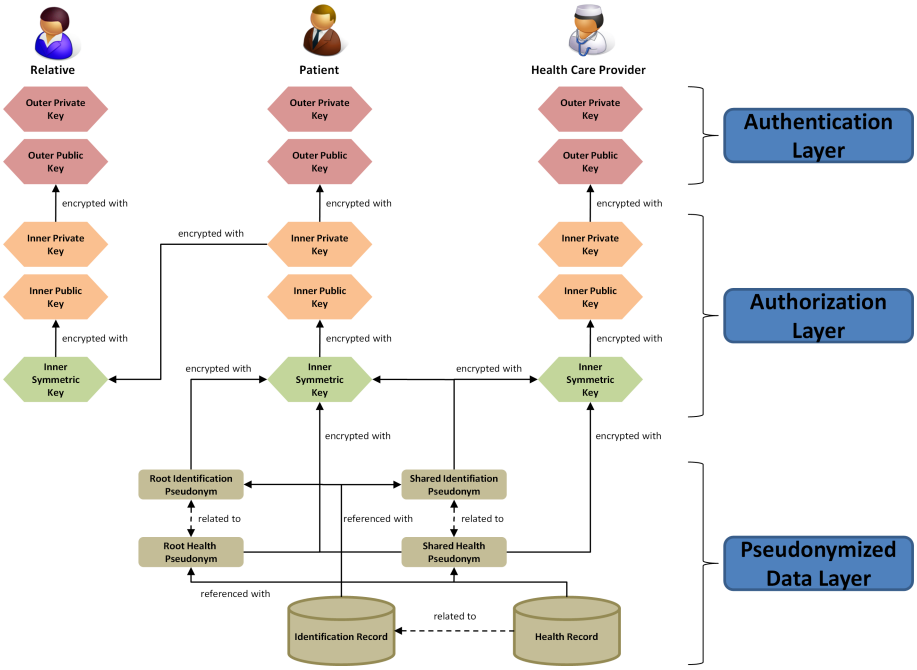

Fig. 3. Layered Security Architecture

arbitrary keywords are ill-suited for range queries and may reveal to much information that could compromise privacy, we allow only structured keywords that are constructed from pre-specified keyword templates. Depending on the actual application domain of the pseudonymization framework, the keyword templates may vary in type and range. For example, for the general e-health scenario, a preferably broad classification of diseases could be used. Standards like the International Statistical Classification of Diseases and Related Health Problems (ICD) or the Logical Observation Identifiers Names and Codes (LOINC) make ideal keyword templates. In addition to these standards, other general purpose templates such as document type (e.g., X-ray image, anamnesis, etc.) and date can complete a specific keyword. Privacy-preservation requires obscuring the relationship between keywords and pseudonyms too, wherefore the keywords' identifiers, like the pseudonyms, are encrypted with the inner symmetric keys and referenced with the encrypted pseudonyms.

3 Pseudonymization and Genetic Data

Like simple anonymization, plain pseudonymization for privacy preservation reaches its limits when genetic data are involved. In recent times, genetic testing has become increasingly popular for identifying genes inducing fatal diseases

(predictive genetic testing) or testing individuals for drug responses (pharmacogenetics) [8]. Especially the results of predictive genetic tests must be handled with great care to prevent discrimination as explained earlier in this article. Predictive genetic testing usually involves the analysis of so-called Single Nucleotide Polymorphisms (SNPs), single-base differences at specific locations in the DNA sequence. Certain SNPs (or combinations of SNPs) are reported to be influential in causing a higher risk for certain illnesses. So by building an individual SNP map and comparing them with known profiles, a higher susceptibility to develop one of the categorized illnesses can thus be identified. Due to the identifying nature of genetic data, depersonalization is often not sufficient to hide the corresponding person: a specific X-ray image of a broken leg usually gets lost in the mass of images stored in a medical database, but depending on the size of a gene sample and the included SNPs (around 1 per 1000 base pairs), this gene string may uniquely identify the corresponding person. There are two ways to address this problem:

- Fragmentation: To reduce the information stored within a single data record, it can be fragmented and the data stored in separate records individually pseudonymized. This is appropriate for predictive genetic test results where usually only a very limited number of SNPs (and corresponding gene sequences) are analyzed. Predictive genetic tests are often issued in packages either suited for a certain group of persons (e.g., age 50+) or to check for a specific group of illnesses (e.g., cardiovascular diseases), involving different points of interests (POIs). By breaking up the individual results, patient-identifying profiling can be prevented.
- Encryption: In clinical research, especially basic research, gene sequences usually represent not single points of interest but rather areas, which means that the genetic data stored include much longer gene strings. Fragmentation could be too tedious here and would require a respectively large number of individual authorizations to cover a complete data set. In this case, full record description is more appropriate where each record is encrypted with its own data encryption key which is also shared when authorized.

For both scenarios, our basic pseudonymization scheme is no longer suitable: (i) Fragmentation requires that the information of the relationships between fragments is kept somewhere; the 1:1 pseudonym relationship is no longer suitable for hierarchical structures. (ii) Encryption of larger document content requires some mechanism to efficiently query within the encrypted record without fully decrypting it. The basic keyword scheme as described in Section 2.3 is no longer adequate to cover these issues.

3.1 Record Description Kit

To handle these new requirements, we developed a separate XML-based record description module or Record Description Kit[4], which allows processing of queries

[4] The Record Description Kit was developed in close collaboration with our colleagues at the Data and Knowledge Engineering Group of the

and updates over encrypted XML documents stored at untrusted servers by exploiting the structural semantics of XML records to return certain parts of the document without decrypting the whole record, and it uses the following mechanisms:

- Schema Labeling: An XML document is defined by a schema definition like an XML Schema or DTD. Each element, attribute, and text node of the XML document is assigned a unique label which encodes structural semantics of the items (path information), resulting in a schema-aware labeling scheme. This labeling schema allows querying for specific parts of the XML document and improves query efficiency because certain query parts can be processed without accessing the database
- Index Structure: In order to speed up frequently executed queries, index structures are created in an XPath-like syntax. These structures support typical queries such as for exact matches, range queries, or queries for structural information like 'all nodes on a specific path'.
- XML Document Storage Structure: XML documents are stored fragmented as key/value pairs, with labels as cryptographically hashed keys and the corresponding XML item as encrypted value. Fragmentation depends on the size of useful pieces of information within the document. Encryption and decryption can be done with standard symmetric cryptographic algorithms (e.g., AES). This ensures that the structure and content of the XML document is hidden from unauthorized viewers. Data stored in this form include index structures, metadata such as schema information, as well as the actual XML documents.

Queries are accepted in the form of (simple) XPath queries and translated to be checked against the labeling schema and index structures for the desired entry. If successful, the corresponding document fragment is retrieved from the database and decrypted.

3.2 Search within (Pseudonymized) Fragmented or Encrypted Records (Genetic Data)

Using the RDK, fragmentation and encryption can be supported as follows:

- Fragmentation: To support record fragmentation without encrypting the health records, the RDK can be used to create an XML-encoded 'inventory' or 'table of contents' listing all records the user is currently allowed to access. Data fragmentation needs to be done at a level such that the individual fragments do not allow patient profiling, but still provide useful information for clinical research. As usual, each fragment is assigned pseudonyms (depending on the number of authorizations) which are also kept in the XML

Johannes Kepler University in Linz, Austria, and is based on their Semantic-based Encrypted XML Document processing architecture SemCrypt (http://www.dke.jku.at/research/projects/semcrypt.html).

document, along with suitable descriptions. Due to their encrypted state, arbitrary and more accurate record descriptions instead of high-level keywords can be stored without compromising confidentiality and privacy.
– Encryption: When fully encrypting the records, the RDK can be used to directly query within the encrypted record for certain parts of the document. Therefore, the record's information needs to be encoded in an XML-conforming data structure, like the Health Level 7 Clinical Document Architecture standard (HL7 CDA). This specification separates the document into a header and body part, where the header's elements are standardized for facilitated data interchange and the body can be arbitrarily defined, depending on the application domain. For genetic data, a specialized HL7 Clinical Genomics (HL7 CG) specification was developed, especially suited to fit the characteristics of genetic information including complete DNA sequences, SNPs, and individual alleles.

Although it seems redundant to pseudonymize fully encrypted records, pseudonymization still has its benefits. Consider the following scenario: Genetic data for clinical research are stored in a central database which is accessible by different research institutions and data contributed by multiple data providers (owners). The documents, which include larger amounts of genetic information, are only disclosed at a need-to-know basis to the research institutions by individual authorizations, and thus data disclosure can be strictly controlled. These authorizations include the encrypted and fragmented XML-encoded document including the labeling schema and index structures, as well as the document-specific cryptographic key required to decrypt the fragments; the dataset is referenced as usual with pseudonyms. While data confidentiality is no issue because of encryption, pseudonymization prevents the individual research institutions from learning which of the records where accesses by fellow researchers, which may be useful in competitive scenarios (e.g., patenting).

4 Formal and Practical Validation

We validated our approach both formally and practically: The formal validation involved the verification of the correctness of the PIPE pseudonymization protocol using the AVISPA tool (Automated Validation of Internet Security Protocols and Applications[5]), while the practical validation was conducted by developing a prototype, which has been implemented in a medium-sized firm offering predictive genetic testing. The AVISPA tool is a protocol verifier which provides implementations of different model-checking techniques in the form of multiple back-ends and requires the security protocol to be defined in the High Level Protocol Specification Language (HLPSL). We modeled each of the core pseudonymization workflows by defining the authorized users (patient, health care provider, relative), the server, and the database as individual roles and assigned them the corresponding knowledge (keys, pseudonyms, etc.). Assuming perfect

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

cryptography and a Dolev-Yao attack scenario, the tool then checked whether an attacker was able to undermine the security goals, i.e, identify the correct pseudonyms and users (identifiers). At any state, the protocol was verified as secure, given that the attacker was provided with publicly available information only. For practical validation, we implemented a prototype to test our approach in a real-life scenario. Using this prototype, the firm pursued to fulfill several goals: (i) to comply with legal requirementsgenetic data need to be stored in an anonymized state; (ii) the data's confidentiality and privacy needs to be protected against attackers (especially insiders); (iii) secured data access should be possible for authorized external personal, and (iv) secondary use for internal statistics and research should be supported. The basic data entities which were decoupled included the patients (i.e., the test orderers) and the test results in the form of SNPs. In this closed scenario, the organization's administration department, managing the complete workflow, requires to know the relationship between test orderers and corresponding test results. Therefore the administration acts as data owner, while the lab has no authorization and thus no knowledge of the test orderers' identities either. We implemented the fragmentation scenario and tested two different cryptographic scenarios: In both scenarios, smart cards were used as authentication tokens. In the first scenario, all cryptographic operations were conducted within the smart cards, but in the second, the main cryptographic operations involved during accessing the authorization and pseudonymized data layer (cf. Figure 2) were executed on the host machines. The resulting performance differences were significant: While the limited calculation power of the smart cards resulted in data retrieval operations lasting several seconds, with host cryptography, the results were produced in well under a second, i.e., with neglectable overhead compared to non-pseudonymized records. Creating specialized index structures (cf. Section 3.1) improved the performance of smart card cryptography by reducing the individual encryption/decryption operations, but as expected never reached the speed of host cryptography.

5 Conclusion

Pseudonymization is a promising technique to fulfill the requirements of data storage and access for primary use as well as privacy-preserving secondary use, but in general requires a sufficiently large number of individuals and records to be effective. We also need to stress the fact that successful pseudonymization (as well as anonymization) requires reliable depersonalization, which can be quite difficult, if not impossible, for certain types of health data. Especially data involving genetic information need to be handled with special care due to the identifying nature. Therefore, we presented a pseudonymization approach that is suitable for pseudonymizing medical records. If required, the basic approach can be extended to handle queryable (selective) document encryption to handle genetic data. In this case, highly sensitive data fragments can be encrypted and still preserve query functionality, while depersonalized and large data, such as

medical images, can be left unencrypted, but are still protected by pseudonymization.

ACKNOWLEDGEMENTS

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.

References

1. Chaudry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., Morton, S.C., Shekelle, P.G.: Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. Annals of Internal Medicine 144(10), 742–752 (2006)
2. Coalition of Genetic Fairness: Faces of genetic discrimination - How genetic discrimination affects real people (July 2004)
3. Congress of the United States of America: Genetic Information Nondiscrimination Act (2008)
4. Council for Responsible Genetics: Genetic discrimination. http://www.councilforresponsiblegenetics.org/pageDocuments/2RSW5M2HJ2.pdf (January 2001)
5. 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)
6. Fischer-Hübner, S.: IT-Security and Privacy: Design and use of privacy-enhancing security mechanisms. Springer, Berlin (2001)
7. Neubauer, T., Heurix, J.: A methodology for the pseudonymization of medical data. International Journal of Medical Informatics 80(3), 190–204 (2011)
8. Roses, A.D.: Pharmacogenetics and the practice of medicine. Nature 405, 857–865 (2000)
9. Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C., Detmer, D.E.: Toward a national framework for the secondary use of health data: An american medical informatics association white paper. Journal of the American Medical Informatics Association 14, 1–9 (2007)
10. Sweeney, L.: k-Anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
11. Thomson, D., Bzdel, L., Golden-Biddle, K., Reay, T., Estabrooks, C.A.: Central questions of anonymization: A case study of secondary use of qualitative data. Forum Qualitative Social Research 6, 29 (2005)
12. United States Department of Health & Human Service: HIPAA Administrative Simplification: Enforcement; Final Rule. Federal Register / Rules and Regulations 71(32) (2006)