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

A methodology for the pseudonymization of medical data

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

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

ARTICLE INFO

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

Keywords:
Computer security
Information management
Electronic medical record
Privacy
Pseudonymization

ABSTRACT

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

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

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

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

[*] *Corresponding author.* Tel.: +43 1 58801 18801.
E-mail address: neubauer@ifs.tuwien.ac.at (T. Neubauer).

1386-5056/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2010.10.016

1. Introduction

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

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

2. Background

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

2.1. *Pseudonymization*

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