Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-751

Recognition and Privacy Preservation of Paper-based Health Records

Stefan FENZ[a,1], Johannes HEURIX[b] and Thomas NEUBAUER [a]
[a] *Vienna University of Technology, Austria*
[b] *SBA Research, Austria*

Abstract. While the digitization of medical data within electronic health records has been introduced in some areas, massive amounts of paper-based health records are still produced on a daily basis. This data has to be stored for decades due to legal reasons but is of no benefit for research organizations, as the unstructured medical data in paper-based health records cannot be efficiently used for clinical studies. This paper presents a system for the recognition and privacy preservation of personal data in paper-based health records with the aim to provide clinical studies with medical data gained from existing paper-based health records.

Keywords. EHR, Security, Privacy, OCR, HL7 CDA

Introduction

The efficient sharing of patient-related data is one of the major prerequisites inE-health to improve the quality of patients' treatment and reduce costs. The introduction of Electronic health records (EHRs) in some areas was a significant step to provide health care providers with a structured and expandable collection of medical data, both for primary and secondary use. However, massive amounts of paper-based health records are still generated and legal obligations demand its storage up to 30 years after the patient's death. Due to the unstructured nature of paper-based health records, this valuable data has no benefit for research organizations and cannot be efficiently used for clinical studies resulting in expensive data acquisition phases and limited samples. Researchers found out that less than one third of the studies actually achieved their resource targets within the originally planned time (cf. [1]). Patients generally agree to provide their medical data for secondary use, but they also express their concerns regarding the application of the personal information for marketing and insurance purposes, disclosure to employers and family members, or the risk of unauthorized access due to security breaches (cf. [2], [3]). The introduction of legal acts, such as HIPAA4, controlling access and disclosure of sensitive health data has highly beneficiary effects for patients' privacy, but also makes research noticeably more difficult (cf. [5]). Therefore, this paper gives an overview of the MEDSEC system, designed for the efficient recognition and pseudonymization of personal data in paper-based health records with the aim to use the resulting data for secondary use.

[1]Corresponding author. Stefan Fenz. Phone: +43(1)5053688, e-mail: fenz@ ifs.tuwien.ac.at.

1. Methods

The MEDSEC system provides clinical studies with pseudonymized and structured medical data gained from existing paper-based health records. The process is divided into four phases:

1.1. OCR

This phase digitizes the content of paper-based health records using Google's open-source OCR engine Tesseract. This engine is one of the most accurate open source OCR engines currently available. This phase not only digitizes the actual content of the health record, it also enriches the OCR output with metadata containing information about the document type (e.g., physician's letter, medical evidence, etc.). With the development of "layout profiles" specific to the health care provider who generates the health records, we could significantly improve the efficiency of the document type classification.

1.2. Annotation.

The Annotation phase automatically recognizes and annotates personally identifiable information (names, addresses, phone numbers, social insurance numbers, etc.). MEDSEC supports the recognition of personally identifiable information (PII) classes as defined in the Health Insurance Portability and Accountability Act (HIPAA). For each information class we defined a set of patterns, rules, look-up tables, and machine learning parameters to correctly recognize and annotate personal information within the paper-based health record. Look-up tables are used for person names, city names, street names, zip codes, organization names, and job titles. Pattern matching is used for information with static structures such as social insurance numbers, phone numbers, e-mail addresses, and IP addresses. Based on the look-up table and pattern matches we designed rules which are used to recognize complex word structures. For example, recognizing a street name and house number as an address or recognizing a misspelled name because of a preceding job title. To enable the recognition of personally identifiable information that is not recognized by the generated patterns, rules, and look-up tables we use machine learning techniques and appropriate training data. We utilized a fine-tuned version of our pattern, rules, and look-up table recognition techniques to generate the training data. By adding the machine learning component to our annotation pipeline we were able to improve the recognition performance on information that is misspelled or was incorrectly recognized in the OCR phase.

1.3. Transformation.

Due to the heterogeneity of clinical IT environments, medical data are often collected and stored in different data formats, which severely complicates data exchange as well as their use in clinical research. The HL7 Clinical Document Architecture (CDA) provides a common standard for representing medical information in a structured way. CDA documents are separated into header and body sections: The highly structured header part is composed of mandatory and optional administrative sections including patient names and addresses, author (i.e., attending physician)

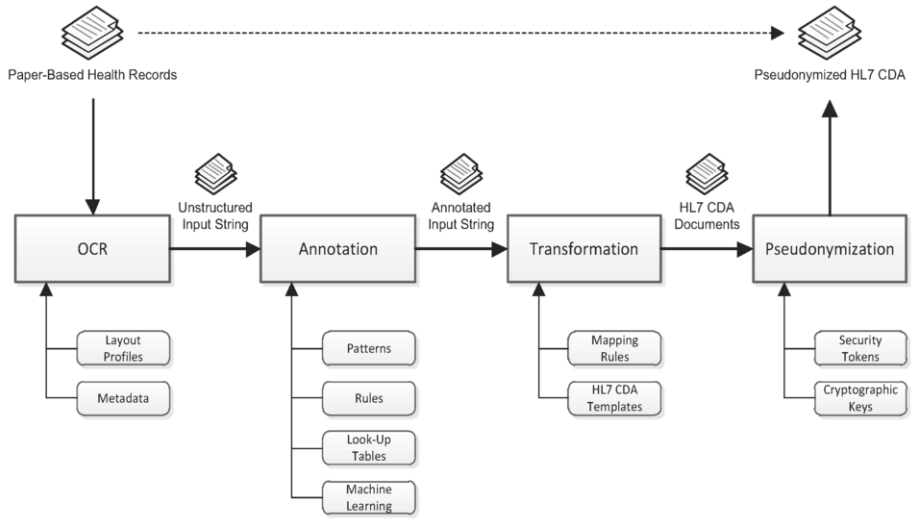

Figure 1. Overview of the MEDSEC architecture

information, or clinical encounters (e.g., in-patient stays). Strict syntactic and semantic requirements ensure automated processing. The body part contains the actual health information which may include highly-structured segments (e.g., ICD codes) as well as arbitrary free-text elements (e.g., discharge summaries). Due to the many different data formats of clinical data, the information needs to be converted into CDA first. MEDSEC uses a rule-based transformation system for automatically identifying relevant information in the input string produced by the Annotation module and transforming it into a CDA-conforming data format (see [6] for a detailed description). It exploits the similarity of medical documents and relies on both structural information (annotations) and text content to identify regions-of-interest which in turn contain the actual contents-of-interest. The latter are then extracted and inserted into document type-specific CDA templates. The rules and CDA templates are expressed in XML to facilitate reuse, while their simple syntax and semantics allow non-technical domain specialists the development of complex rules without the need to master complicated formalisms.

1.4. Pseudonymization.

In order to guarantee patient's privacy, e.g. according to HIPAA, MEDSEC makes use of data pseudonymization (see [7] for a detailed description). Similar to anonymization, with pseudonymization the actual health records are stored in cleartext while the corresponding PII is kept secret. But in contrast to anonymization, PII is not simply removed but replaced by a pseudonym such that the patient cannot be linked to their corresponding health records without knowing a certain secret, i.e., the link is protected by encryption where this secret corresponds to the decryption key. Thus, this secret allows relinking of patient to medical records only under strictly defined circumstances, whereas the medical records are stored in an anonymized state. The architecture of the MEDSEC pseudonymization module consists of three layers. The authentication layer demands the user to prove his identity. The authorization layer allows the user to share his data with other users, such as relatives or health care providers and the

peudonymized data layer harbors the actual data together with the pseudonyms. The cryptographic operations (including encryption and decryption) are basically carried out on the server (by using a HSM). The authentication procedure that is based on a challenge/response procedure makes use of the user's security token (e.g., smartcard) taking the role of (i) a secure keystore for the user's private key and his authentication credentials and (ii) a client-side cryptographic module.

2. Discussion

MEDSEC introduces new methods for the automated identification of personal and medical data in paper-based health records. It provides common synonyms and formal specifications of personal and medical data elements to enable their automated detection in paper-based health records. Since the sole digitization of paper-based health records is not sufficient for providing clinical research with appropriate data, MEDSEC enriches the digitized data with standard-compliant metadata (e.g., according to the HL7 standard). The concept of pseudonymization use within MEDSEC allows the data to be associated with a patient only under specified and controlled circumstances and, thus, mitigates the security shortcomings of existing approaches.
MEDSEC was implemented into a software solution and tested within a national healthcare provider in Austria that treats about 250.000 inpatients and 600.000 outpatients annually. The evaluation results showed that MEDSEC simplifies the analysis of medical data by providing more representative samples. Thereby, it reduces the time required for carrying out clinical research (including clinical trials) because clinical studies can be carried out faster. Additionally, the larger sample results in more reliable and significant outcomes. Digitized health records reduce costs for hospitals and research organizations by saving expensive archive space of paper-based health records and allowing the categorization of data and, thus, the fast and efficient search for specific information and the transformation of medical data into HL7 allows the more efficient administration and use of this data in clinical environments.

Acknowledgments. The research was funded by BRIDGE (#824884) and by COMET K1, FFG – Austrian Research Promotion Agency.

References

[1] Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, Entwistle V, Garcia J, Roberts I, Grant A, Grant A; STEPS group. Recruitment to randomised trials: strategies for trial enrolment and participation study. The steps study. Health Technology Assessment 2007; 11(48): 9-105.
[2] Willison DJ, Keshavjee K, Nair K, Goldsmith C,. Holbrook AM. Patients' consent preferences for research uses of information in electronic medical records: Interview and survey data. British Medical Journal 2003; 326: 373
[3] Simon SR., Evans JS, Benjamin A, Delano D, Bates DW. Patients' attitudes toward electronic health information exchange: Qualitative study. Journal of Medical Internet Research 2009; 11(3): e30
[4] United States Department of Health & Human Service. HIPAA Administrative Simplification: Enforcement; Final Rule. Federal Register / Rules and Regulations 2006; 71(32): 8389-8433
[5] Ness RB. Influence of the HIPAA privacy rule on health research. Journal of the American Medical Association 2007; 298(18): 2164-2170

[6] Heurix J, Rella A, Fenz S, Neubauer T. Automated transformation of Semi-Strucutred Text Elements. in Americas' Conference on Information Systems (AMCIS); 2012.
[7] Neubauer T, Heurix J. A methodology for the pseudonymization of medical data. International Journal of Medical Informatics 2011; vol. 80, no. 3, pp. 190-204, 2011.
[8] W3C. Platform for Privacy Preferences (P3P) Project; 2007
[9] Squicciarini A, Bertino E, Ferrari E, Ray I. Achieving privacy in trust negotiations with an ontology-based approach,. IEEE Transactions on Dependable and Secure Computing 2006; 3:13-30
[10] Waegemann CP. Status Report 2002: Electronic Health Records; 2004.
[11] Anton AI, Earp JB, Reese A. Analyzing Website privacy requirements using a privacy goal taxonomy. in Proceedings od the IEEE Joint International Conference on Requirements Engineering, 2002; 23-31.