Journal of Biomedical Informatics 77 (2018) 81–90

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Controlled searching in reversibly de-identified medical imaging archives

Jorge Miguel Silva[a,*], Eduardo Pinho[a], Eriksson Monteiro[b], João Figueira Silva[a], Carlos Costa[a]

[a] DETI/IEETA, University of Aveiro, Portugal
[b] BMD Software, Portugal

ARTICLE INFO

Keywords:
Medical Imaging
PACS
DICOM
Medical databases indexing
De-identification
Pseudonymisation

ABSTRACT

Nowadays, digital medical imaging in healthcare has become a fundamental tool for medical diagnosis. This growth has been accompanied by the development of technologies and standards, such as the DICOM standard and PACS. This environment led to the creation of collaborative projects where there is a need to share medical data between different institutions for research and educational purposes. In this context, it is necessary to maintain patient data privacy and provide an easy and secure mechanism for authorized personnel access. This paper presents a solution that fully de-identifies standard medical imaging objects, including metadata and pixel data, providing at the same time a reversible de-identifier mechanism that retains search capabilities from the original data. The last feature is important in some scenarios, for instance, in collaborative platforms where data is anonymized when shared with the community but searchable for data custodians or authorized entities. The solution was integrated into an open source PACS archive and validated in a multidisciplinary collaborative scenario.

1. Introduction

Over the last years, the use of digital medical imaging in healthcare institutions has significantly increased and the expectation is for it to continue. For instance, Frost & Sullivan registered that the US in 2005 had a cumulative storage of 119,325 Terabytes and in 2015 reached 300,000 Terabytes [1]. It was also estimated that data production would be over 1 Exabyte in 2016 [2] and that storage volume will double every 24 months [3].

This enormous production of data was followed by the development of new digital medical imaging systems, which become fundamental in clinical practice to support diagnosis and treatment.

Research and industry are constantly updating and proposing new systems, namely imaging acquisition modalities, grounded in the universally accepted concept of Picture Archiving and Communication System (PACS). This concept includes hardware and software for imaging acquisition, storage, distribution and visualization. It has revolutionized medical practice during the last two decades and its digital nature has brought many benefits, including facilitated image manipulation and interpretation for value-added diagnosis. Furthermore, it provides support for improved and advanced workflows, which triggered speedier healthcare delivery and a reduction in operational costs [4,5].

The normalization of data format and network services in PACS environments arrives with the development and adoption of the Digital Imaging and Communications in Medicine (DICOM) standard, fulfilling the essential requirements for systems' interoperability and accurate representations of medical imaging data. This standard was designed to cover all functional aspects of digital medical imaging laboratories, including data structures and codification, but also data discovery and retrieval services [6]. A DICOM object contemplates instance representation (e.g. image pixel data) but also associated meta-data organized in groups of related attributes. However, traditional archive databases only store a small number of attributes imposed by the DICOM Information Model (DIM). So the majority of information elements contained in objects' metadata are not searchable in those systems. In some recent platforms, this issue was solved with the inclusion of document-oriented retrieval mechanisms [7].

Independently of the repository database approach, maintaining privacy and confidentiality of the patient's personal health information (PHI) is a key point in any implementation, even more when many consider health information to be among the most confidential types of personal information [8]. This issue is particularly critical in today's environments where storage outsourcing, data sharing and collaborative platforms are a reality. Hence, protecting patient privacy is extremely important for any electronic healthcare records (EHR) system and legal protection rules must be enforced [9], with particular focus on ensuring the confidentiality, integrity and availability (CIA) of health information [8]. To address the challenging issue of preserving medical data privacy, various techniques can be implemented including

[*] Corresponding author.
E-mail address: jorge.miguel.ferreira.silva@ua.pt (J.M. Silva).

https://doi.org/10.1016/j.jbi.2017.12.002
Received 24 July 2017; Received in revised form 22 November 2017; Accepted 4 December 2017
Available online 07 December 2017

cryptography, access control and data anonymization [10].

DICOM meta-data includes identifiable information about the patient, the study, acquisition equipment, clinical staff, etc. Moreover, part of this information can also be burned into the pixel data in some medical imaging modalities, such as ultrasound. As such, managing such sensitive data demands proper protection to ensure patient data privacy, especially in collaborative scenarios.

Medical imaging anonymization is not a recent subject in the literature. However, there are no reports of DICOM medical imaging repositories that provide anonymization of private information present in the pixel data and meta-data, ensuring at the same time search functionalities that allow authorized users to identify the patient studied or use identifiable attributes in queries. Ordinary platform users only need to have access to anonymized data, preventing Cyphertext-only attacks where the distribution of the anonymized data is known as well as the language in which the information is written [11,12]. In order to control a given entity's access to health information, mechanisms based on Role-Based Access Control (RBAC) or Situation-Based Access Control (SitBAC) can be used [13]. The deployment of such systems, with granular control over what health data is disclosed for each entity layer [14], is particularly useful in multi-disciplinary collaborative scenarios, for instance, in research platforms or screening programs where images are shared between distinct specialists, including computer science researchers and doctors. In these scenarios, it is usual to signal certain cases for subsequent analysis by authorized personnel, such as specialists performing blind diagnosis or annotation procedures. Sometimes, they signal the case so that authorized personnel can reverse the anonymization process and retrieve complementary information, thus making a more accurate diagnosis [15]. Another usage scenario occurs when statistical studies are being carried out, and there is a need to analyze factors that are related to the manifestation of the disease. These factors often consist of patients' private data such as age and gender, and as such there is the need to de-anonymize this information [16].

This paper presents a solution to support the previously described scenarios. It comprises a pipeline that first applies an automatic visual anonymizer, which uses the image meta-data as well as a machine learning model, removing only the patient's private information and preserving other textual information burned into the image. Then, a reversible anonymizer of medical imaging data is used that maintains search capabilities over the original DICOM data elements. This pipeline was integrated into an open source PACS archive to contemplate the multidisciplinary collaborative platform scenario.

2. Related work

A DICOM persistent object is composed of data elements (or attributes) that are represented by a unique tag with specific values and data types [17]. These meta-data often possess identifiable information about the patient, study, institution, etc. There are two known methods for de-identification of patient-related information: *anonymization* and *pseudonymization*. The first method removes information carried by header elements or replaces the information with random data such that the remaining information cannot be used to reveal the patient's identity at all. The second method replaces most identifying fields within a data record using one or more artificial identifiers that could be used by authorized personnel to track down the real identity of the patient. This is the most appreciated method since clinical practice benefits from this kind of tracking, namely when additional findings occur in this situation [18].

Besides meta-data, patients' personal information may be present in the image pixel data. This happens mostly in modalities that existed before the era of digital imaging, and the process ensured a secure way of associating an examination with the respective patient. With the transition to DICOM, some of these modality devices remained with this modus operandi. Among others, the majority of these devices belong to the Ultrasound modality (US) and exams that make use of the External-camera Photography modality (XC) exhibit this behavior as well. Those identification elements burned in the image must be removed to ensure an anonymization procedure compliant with HIPAA [19].

There are several tools to perform de-identification of meta-data and pixel data in order to fulfill the requirements of patient data protection. However, to the best of our knowledge, there is currently no anonymized repository system able to provide a controlled search mechanism over the original DICOM data elements. In other words, none of the anonymization tools provide free text search over pseudoanonymized content and respective de-anonymized search results when the user is authorized to access that information in the PACS.

With regard to meta-data anonymization, DICOM Standard for medical imaging communications and storage defines an attribute level confidentiality mechanism [20]. Nevertheless, this mechanism may not have been widely adopted yet [21]. Furthermore, the de-identification process will depend on the specific legal framework of each country and on particular circumstances of the use case.

Related to meta-data anonymization, Aryanto et al. [18] evaluated the ability of open source tools to remove a patient's PHI from DICOM headers. The selection criteria for the tools was the frameworks' ability to perform de-identification and availability as freeware or an open source tool. The tools selected were Conquest DICOM software,[1] RSNA Clinical Trial Processor (CTP),[2] DICOM library,[3] DICOMworks,[4] DVTK DICOM anonymizer,[5] GDCM,[6] K-Pacs,[7] PixelMed DICOMCleaner,[8] Tudordicom,[9] and YAKAMI DICOM tools.[10] The de-identification was performed to fifty header elements. The results showed that only five tools could properly de-identify the defined DICOM elements, and in four cases, only after careful customization.

Onken et al. [15] have shown that reversible de-identification of DICOM data can be achieved with good coverage by generating anonymization policies from the DICOM standard. The solution supposes that only DICOM searching services are used, which only supports a limited number of mandatory DIM attributes.

Regarding annotations in pixel data, there are open source tools that can assist the manual removal of burned pixel information [22]. On the other hand, there are also automated processes that use commercial optical character recognition (OCR) solutions for detecting burned-in text on the image [23]. The problem with the latter solution is that it is not possible to remove only protected health information annotations and preserve other important textual information such as measurements. Huang et al. [24] also proposed a system for the de-identification of medical images, using pre-defined area filters and OCR to recognize and remove patient identifiers. While the authors reported a 65% success rate for pixel data de-identification, pixel data de-identification was not reliable due to errors when recognizing ambiguous characters and due to the medical image background.

The anonymized medical imaging repository proposed in this article combines a solution for removing protected health information from pixel data, based on a machine-learning model that provides reliable accuracy, and a novel reversible de-identifier of meta-data that retains search capabilities from the original DICOM data index.

[1] Conquest DICOM: http://www.medfloss.org/node/93.
[2] CTP: http://www.rsna.org/ctp.aspx.
[3] DICOM library: http://www.dicomlibrary.com/.
[4] DICOMworks: http://www.dicomworks.com/.
[5] DVTK DICOM anonymizer: http://www.dvtk.org/.
[6] GDCM: github.com/malaterre/GDCM.
[7] K-Pacs: http://www.k-pacs.net/.
[8] PixelMed DICOMCleaner: http://www.pixelmed.com/cleaner.html.
[9] Tudordicom: http://www.santec.tudor.lu/project/dicom.
[10] YAKAMI DICOM tools: http://www.kuhp.kyoto-u.ac.jp/.