Desensitization and Cleaning Technology of CT Image of Parotid Tumor based on DICOM

Jiantin Yuan\textsuperscript{1,*}, Wenqiang Zhang\textsuperscript{1}, Wei Wu\textsuperscript{1}, Xiaoyi Lv\textsuperscript{1}, Xue Bai\textsuperscript{2}, Chen Chen\textsuperscript{2} and Yue Hong\textsuperscript{3}

\textsuperscript{1}School of Software, Xinjiang University, Xinjiang Urumqi 830046, China
\textsuperscript{2}School of Information Science and Engineering, Xinjiang University, Xinjiang Urumqi 830046, China
\textsuperscript{3}Hospital of Radiological Imaging Center, Xinjiang Uygur Autonomous Region People's Hospital, Xinjiang Urumqi 830001, China

*E-mail: steven@xju.edu.cn

Abstract. CT image data are widely used in model training and data mining in the fields of medical image processing and medical big data analysis. Sensitive information on the data set needs to be desensitized before use. Traditional CT image data desensitization often wipe Remove most of the patient’s identification information. The data has not been cleaned, and the quality and analytical value of the data can’t be guaranteed after desensitization. In order to solve the privacy protection and data quality problems of DICOM-based image datasets during use, this paper uses symmetric encryption, numerical transformation and invalidation desensitization strategies to desensitize CT images of parotid tumor CT images based on DICOM. The desensitized encrypted data are decrypted to determine the true identity of the patient for data cleaning. Data cleaning is performed on the desensitized and decrypted data according to the formulated process. While meeting the desensitization standard to protect the privacy and safety of patients, the method of this paper greatly guarantees the quality, statistical characteristics and analytical value of the desensitized data set, and provide a feasible solution to the safe use and quality control of CT image data sets in the future.

1. Introduction

With the rapid spread of medical information and the arrival of the era of artificial intelligence [1], medical image data are widely used in the field of medical image processing and medical big data analysis. both of them need huge medical image data onto model training and data analysis [2]. In the medical image data, the CT image accounts for a large proportion based on DICOM standard, which contains a large number of sensitive information on patients, such as name, Patient ID, date of birth, diagnostic information and so on[3]. Sharing or publishing the original data directly without processing will inevitably disclose patient’s sensitive information and cause great trouble to patients, so it is particularly important to desensitize CT images [4]. However, the protection for patient privacy data and the analytical value of its data are two contradictory conflicts [5]. For the desensitization of traditional CT image data, it is often necessary to erase most of the identification information about patients, so the analytical value of the data is greatly reduced [6]. If the patient privacy data information is completely retained, although the mining value of the data can be maximized, the risk of patient privacy disclosure can’t be controlled [7]. Therefore, in the formulation of desensitization
strategy, it is necessary to protect the privacy of patients while maximizing the value of data analysis [8].

At present, the desensitization method of CT image is relatively simple based on DICOM, which greatly destroys the potential value of the data after desensitization. Substandard, unreadable or repetitive data may be produced in the desensitization process, which can’t guarantee the quality of the desensitized data set [9]. Therefore, in order to consistent with the standards after the desensitization, data cleaning is indispensable [10]. In this paper, the patient unique identifier Patient ID in the CT image of parotid tumor based on DICOM is processed by a recoverable symmetric encryption algorithm [11], other data are desensitized by invalidation and numerical transformation, and a data cleaning process is developed for the desensitized data set [12]. After data desensitization and encryption, the Patient ID can be decrypted, and the true identity of the patient can be determined according to Patient ID, which can be used to assist the data cleaning work to determine whether there is a non-unique or non-desensitized CT image of parotid tumor. In this paper, the CT image desensitization and cleaning method of parotid tumor based on DICOM not only effectively protects the privacy and security of patients, but also preserves the availability of data to a large extent, ensures that the value of data is not destroyed, and improves the quality and application range of the data set [13]. It will play an important role in the security and quality control of medical image data [14].

2. Experiment

2.1 Introduction of experimental methods

Medical digital image and communication is also called DICOM standard. The latest 3.0 Universal version of DICOM standard unifies the interface standard of medical image and its related data information on different systems and application environments, basically with “.dcm” as the suffix. Its content contains most of the commonly used data in medical scenes, which can reflect medical equipment, image data and patient personal information[15]. Therefore, the medical image information on DICOM standard plays a very important role in the digital networking of medical information system and the development of medical big data, and is widely used in the field of medical image processing of artificial intelligence and big data mining [16].

Pydicom is a pure Python package that handles DICOM files. It can obtain and change the data of DICOM standard in an easy to use way, and it can also save the changed data as a new DICOM type file [17]. Pydicom can run on any platform under the Python interpreter, requiring little configuration other than the installation of the Numpy module. Pydicom package is currently the mainstream medical image processing package in Python.

Data cleaning refers to the process of processing datasets according to data cleaning rules to obtain standard, complete and high-quality data. [18]. Based on DICOM standard, CT image contains medical information and image information in a specific format, which involves many ethical issues of patient privacy and privacy protection. Before data cleaning, it is necessary to desensitize patient’s private information, so it is necessary to develop appropriate data desensitization and cleaning schemes [19].

2.2 Desensitization rules and cleaning plan

Data desensitization under the condition of retaining the original characteristics of the data, some sensitive data are processed by desensitization rules to ensure the reliable protection of users' private data. Data desensitization will cause some information loss and affect the results of analysis, so the desensitization strategy needs to make a balance and trade-off between data confidentiality and analysis ability. In order to increase the sharing goal and use scope of the original data set on the premise of ensuring the privacy and security of patients. In this paper, an appropriate desensitization strategy is developed according to the application scene of parotid tumor CT image based on DICOM. The schematic diagram of the desensitization strategy is shown in Figure 1.
3) Schematic diagram of desensitization strategy

1) Symmetric Encryption: a reversible desensitization method. Encryption keys and encryption algorithms are used to encrypt sensitive data. The encrypted data is the same as the sensitive data in terms of logical rules, and the desensitized data can be restored to the original data after decryption.

2) Invalidation: the use of a special symbol (such as Null or *** ) to replace the original sensitive data or part of the sensitive data.

3) Replacement: replace all or part of the sensitive information in the original data with imaginary data. The Replacement is usually irreversible, resulting in the loss of part of the information value of the data.

4) Numerical transformation: refers to the controllable and uniform adjustment of numerical and date-type raw data through specific functions in order to complete the desensitization of specific values while retaining some of the statistical characteristics of its original data, effectively controlling desensitization The authenticity and statistical characteristics of the data.

Data cleaning data analysis and application process is an essential process. Based on the CT image of parotid gland tumor of DICOM, a data cleaning process was developed to remove some data that did not meet the standard, did not desensitize or repeatedly appeared after desensitization. The specific process includes the following steps: ① desensitizing and encrypting the CT image of parotid gland tumor; ② decrypting the encrypted CT image of parotid gland tumor to recover the patient's Patient ID; ③ determining the data cleaning rules for the desensitized and decrypted CT image of parotid gland; ④ defining and checking the cleaning process according to the desensitized field information and cleaning rules; ⑤ inputting the encrypted and decrypted parotid gland The CT image of tumor was cleaned; ⑥ the cleaned data were reviewed and saved.

The data cleaning technology after desensitization of CT image of parotid tumor should not only consider the basic requirements of traditional data cleaning technology, such as data consistency check, abnormal data screening and duplicate data removal, but also consider the unified rules of integrity, uniqueness, readability and security of CT image of parotid tumor to determine whether the information of Patient ID and gender in the image before and after desensitization is consistent, confirm whether the patient's name, date of birth, organization name, equipment name and other information are desensitized correctly.

3. Experimental results and results analysis

3.1 Experimental environment and data set
In this paper, CT images of parotid tumors in DICOM3.0 format obtained from the People's Hospital of Xinjiang Uygur Autonomous Region were used as the original data set. Desensitize the sensitive data according to the established desensitization rules, and decrypt the Patient ID information of the
CT image of the desensitized and encrypted parotid gland tumor after desensitization to assist data cleaning. The CT image of parotid gland tumor after desensitization and decryption was applied to data cleaning. The experimental environment is shown in Table 1.

| Language | Python |
|----------|--------|
| System   | Microsoft Windows 10 X64 |
| Development platform | PyCharm, Pydicom Development kit, Re module |
| Features | Object-Oriented, Cross-Platform, Portable Extensions |
| Application | AI, Big Data, Cloud Computing, Digital Image Processing |

### 3.2 Overall process framework
The overall program flow can be divided into two parts: data desensitization and data cleaning: ① The function of data desensitization is to use the established desensitization strategy to complete the desensitization of the data of the original data set ② The function of data cleaning is based on the established The data cleaning process performs data cleaning on the desensitized and decrypted data set. The overall process of image desensitization and cleaning is shown in Figure 2.

**Figure 2. Overall process of image desensitization and cleaning**

### 3.3 Specific case realization
According to the established desensitization strategy and data cleaning process, the function of this program is to complete the desensitization and cleaning technology of parotid tumor CT image data set based on DICOM. This paper chooses PyCharm as the development platform, uses the method in medical image processing package Pydicom (pydicom.read_file) to read the data set information, uses RE module and symmetrical encryption algorithm to complete data desensitization, and then cleans the decrypted data set. The specific implementation of the desensitization strategy is as follows:

1) Patient ID is the unique identifier of the patient, which can be used to confirm the identity of the
patient. Consider using a reversible desensitization algorithm for encryption. This paper uses a symmetric encryption algorithm including key, and uses key to encrypt or decrypt Patient ID. The encrypted CT image of parotid tumor can be used by a third party, and the real information of the patient will not be exposed in the process of use. On the other hand, the internal system of the hospital or the doctor can decrypt the original Patient ID, through the key and encryption algorithm to confirm the true identity of the patient.

2) In the application process of CT images of parotid tumors, data may be divided according to age and gender, so gender information is retained, and part of the true data is retained by numerical transformation on the date of birth. For example, desensitization of 19950513 to 19900101.

3) Partial invalidation strategy is adopted for the name, and the rest of the Patient Name's last name is replaced by ***. In order to facilitate the group management and differentiation of the data, the information such as the name of the institution and the name of the equipment are unified and standardized by means of replacement. For data cleaning, the flowchart is shown in Figure 3.

![Figure 3. Data cleaning process](image)

Data comparison; open the data set processed by the desensitization and decryption program, and confirm that the number of parotid tumor CT images in the desensitized data set is consistent with the original data to ensure the integrity of the data set.

Desensitization data cleaning; determine the cleaning method according to the overall flow chart of image desensitization and cleaning, and use the cleaning strategy to clean the desensitized data set to confirm that the CT image of the parotid tumor complies with the DICOM standard to ensure the readability of its content. Remove unreadable, non-standard, and missing data.

Data check; check whether the sensitive data of the CT image of the parotid tumor is desensitized correctly according to the desensitization strategy. Determine whether the Patient ID after the image decryption is the same as the original data, and delete the data that is not correctly desensitized or the Patient ID is incorrect.

Data retention; high-quality and effective data after data desensitization and data cleaning are retained to facilitate subsequent data analysis and application.

![Figure 4. CT image of the original parotid tumor](image)
3.4 Result analysis
In order to demonstrate the effectiveness of the technique of desensitization and cleaning of CT images of parotid tumors based on DICOM in this paper, desensitization and data cleaning of CT images of parotid tumors were performed. The CT image of the original parotid tumor is shown in Figure 4. In this paper, Figure 4 is desensitized to Figure 5, where the data marked with yellow line is the Patient ID after symmetric encryption, Use the invalidation method to process the name, and replace the part with ***. The data in the red box is the result of invalidation and desensitization by numerical transformation. The encrypted data (Patient ID) in Figure 5 can be decrypted on the application program, and the decrypted image is shown in Figure 6. The part marked in the red box in the figure is the real Patient ID after decryption, which can effectively assist the subsequent data cleaning work.

Figure 5. CT image of parotid tumor after desensitization and encryption

Figure 6. CT image of parotid tumor after desensitization and decryption

4. Conclusion
The rapid development of information technology and Internet industry has promoted the strong demand for datasets from artificial intelligence and big medical data. At the same time, privacy protection and data quality problems have arisen. In order to ensure the privacy security and data quality of CT image data set of parotid tumor based on DICOM, in this paper, the CT image of parotid gland tumor based on DICOM standard is desensitized by symmetric encryption, partial masking and numerical transformation. The CT image of the desensitized parotid gland tumor meets the desensitization standard. To avoid omission during desensitization, CT images of unsensitized or unavailable parotid tumors should be avoided during application. The desensitization data needs to be cleaned. The CT images of the desensitized parotid gland tumors were screened by data cleaning to remove the incomplete and data-lost images. In this process, the real identity of the patient can be determined by decrypting the desensitization encrypted data and obtaining the Patient ID of the original image as identification for subsequent data cleaning. In this paper, the CT image data set of the desensitized parotid gland tumor is cleaned to ensure the quality of the data set. In the future, this method will play an important role in the safe use and quality control of medical data sets.

Acknowledgement
This work was supported by the Science and Technology Project on aid to Xinjiang Uygur Autonomous Region "Development and Application Demonstration of Medical Big Data Increment
Desensitization Platform Based on Mining” (No.2019E0215).

References
[1] Chen Tian-ying, Chen Jian-feng. Intelligent Data Masking System for Big Data Productive Environment[J]. Communications Technology, 2016, 49(07): 915-922.
[2] Morris M A, Saboury B, Burkett B, et al. Reinventing radiology: big data and the future of medical imaging[J]. Journal of thoracic imaging, 2018, 33(1): 4-16.
[3] Price W N, Cohen I G. Privacy in the age of medical big data [J]. Nat Med, 2019, 25(1): 37-43.
[4] Wang Yang, Liu Li-bo. Research and Implementation of Desensitization System for CT Medical Images Based on DICOM[J]. Modern Computer, 2019, 72-75.
[5] Al Hamid H A, Rahman S M M, Hossain M S, et al. A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using Fog Computing Facility With Pairing-Based Cryptography[J]. IEEE Access, 2017: 22313-22328.
[6] Lee C H, Yoon H J. Medical big data: promise and challenges[J]. Kidney research and clinical practice, 2017, 36(1): 3.
[7] Cui B J, Zhang B H, Wang K Y. A Data Masking Scheme for Sensitive Big Data based on Format-Preserving Encryption[C]//2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). IEEE, 2017, 1: 518-524.
[8] Boukobza E. System and method for data masking: U.S. Patent 8,826,370[P]. 2014-9-2.
[9] Hoffman S. Medical big data and big data quality problems[J]. Conn. Ins. LJ, 2014, 21: 289.
[10] Kulkarni P S, Bakal J W. Survey on data cleaning[J]. International Journal of, 2014.
[11] Al-Haj A. Providing integrity, authenticity, and confidentiality for header and pixel data of DICOM images[J]. Journal of digital imaging, 2015, 28(2): 179-187.
[12] Mansfield-Divine S. Masking sensitive data[J]. Network Security, 2014, 2014(10): 17-20.
[13] Kayhart D. Using an Existing DICOM Infrastructure to Enhance the Availability, Quality, and Efficiency of Imaging Throughout the Healthcare Enterprise [J]. J Digit Imaging, 2019, 32(1): 75-80.
[14] Andreu-Perez J, Poon C C Y, Merrifield R D, et al. Big data for health[J]. IEEE journal of biomedical and health informatics, 2015, 19(4): 1193-1208.
[15] Shin H B, Sheen H, Lee H Y, et al. Digital Imaging and Communications in Medicine (DICOM) information conversion procedure for SUV calculation of PET scanners with different DICOM header information [J]. Phys Medica, 2017: 243-248.
[16] Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era[J]. Data science journal, 2015, 14.
[17] Litviev D S, Trivedi H, Panahiaza M, et al. Automatic Labeling of Special Diagnostic Mammography Views from Images and DICOM Headers[J]. Digit Imaging, 2019, 32(2): 228-233.
[18] Li Z, Sun L, Higgs R. Research on, and Development of, Data Extraction and Data Cleaning Technology Based on the Internet of Things[C]//2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). IEEE, 2017, 2: 332-341.
[19] Hao Ye, Tang Qiaohong, Li Jiage, et al. Study on Data Cleaning Technology in DICOM format Medical Image Quality Control[J]. China Medical Devices, 2018, 33(12): 10-13.