Design and Implementation of a Desensitization Method of CT Medical Image based on DICOM

Jiantin Yuan1*, Wei Wu1, Wenqiang Zhang1, Xiaoyi Lv1, Chen Chen2

1School of Software, Xinjiang University, Xinjiang Urumqi 830046, China
2School of Information Science and Engineering, Xinjiang University, Xinjiang Urumqi 830046, China

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

Abstract. Medical big data, medical image processing and other fields are based on a large number of clinical data. Data desensitization is of great significance to the privacy protection of patients. CT image desensitization based on DICOM standard protocol only uses the strategy of hiding or deleting the data on the image, which greatly weakens the value of data use and cannot meet the normal value demand. Data desensitization should consider that it can still meet the needs of data mining and other applications on the basis of ensuring the privacy security of users. Based on the above application requirements, a strategy to anonymize patient identity is developed. Based on this strategy, a desensitization method that can meet the actual desensitization standards and analysis application requirements at the same time is proposed. This method can not only protect the privacy of patients to a great extent, but also ensure the availability and authenticity of desensitized data. In the future, this method has great potential in the application of CT image desensitization.

1. Introduction
The rapid development of health care big data, medical image processing and other fields has promoted the progress of medical and health care in our country, but the privacy of patients has also appeared security risks [1]. The research and development of health care big data, medical image processing and artificial intelligence products are supported by massive medical image data, among which CT image based on DICOM standard is an important source [2]. Since the image contains a lot of real information about the patient, it needs to be desensitized before research and application. The analysis and application of medical image data can bring great value [3]. On the basis of ensuring the privacy and security of patients, how to meet the needs of medical big data value mining [4] is also a problem that must be considered for data desensitization.

The usual CT image desensitization method based on DICOM standard protocol only uses the strategy of masking or deleting to desensitize the data on the CT image [5], and the data on the image is masked or not displayed. Although it achieves the purpose of desensitization, it also loses the value of analysis and application. For example, the diagnostic information on CT images can be classified and managed by users according to the location or time of diagnosis. Therefore, it is of great significance to study a method which can not only reach the desensitization standard, but also make the desensitized data still have a certain application value. By analyzing the data attributes of patients in CT images and dealing with the identity attributes and quasi-identification attributes in patient data, a strategy of keeping patient identity anonymity is developed. According to this strategy, a CT image
A desensitization method based on DICOM standard is developed by using Python language. This method is of great significance in ensuring the privacy and security of patients in the application of medical big data and improving the operational efficiency of medical data management in radiology department.

2. Related work

2.1. DICOM standard and CT image brief introduction
DICOM standard (Digital Imaging and Communications in Medicine) provides a consistent interface for various medical imaging products [6], which greatly simplifies the exchange of medical image information [7], such as medical imaging information system (PACS) and remote radiology system [8]. The suffix of the medical image DICOM file is .dcm. The medical image information and related text information are stored on the file [9], and the patient's personal basic information and medical diagnostic information are all in it [10]. Private information such as the patient's name, gender and date of birth will also be displayed on the CT image in DICOM format, as well as the ID, location, time and type of the examination. At present, the analysis and application of data on medical images based on DICOM standard is very common, which provides rich data resources for health care big data [11].

CT (Computed Tomography) continuously scans a certain part of the human body, also known as computed tomography [12]. The image of CT is clear and the scanning time is fast, so it can be used to detect a variety of diseases. CT provides a factual basis for the diagnosis of tumors [13], and is of great significance for the diagnosis of central nervous system diseases, head and neck diseases, chest, abdomen and other diseases [14].

2.2. Data desensitization overview
Data desensitization, also known as data deformation, data bleaching or data desensitization, some sensitive data can be processed accordingly through the desensitization strategy, so as to realize the protection of sensitive private data [15]. Sensitive information usually involves the privacy of individuals or groups, such as name, age, ID number, diagnosis, organization name, account password, etc. Once these private information is disclosed, it will cause great trouble to individuals or enterprises [16]. Desensitization strategy should be determined according to the actual specific needs, try to ensure the security of the data and make the desensitized data available [17].

| Desensitization method names | Desensitization method description |
|------------------------------|-----------------------------------|
| Replacement                  | Replacement refers to replacing all the original sensitive data with prepared data to achieve the effect of disguise. The replacement data cannot be restored to the original data. |
| Encryption                   | Encryption refers to the use of certain encryption rules to process sensitive data, so that sensitive data loses its original characteristics. |
| Masking                      | Masking is the use of meaningless identifiers to replace sensitive data in whole or in part, making it impossible for outsiders to access the original data. |
| Deletion                     | Deletion refers to setting sensitive data to a null value or removing it, making sensitive data invisible. |
| Change                       | Change refers to adjusting specific types of data in accordance with established rules to achieve the effect of disguise. The transformed data still retains the original statistical characteristics. |
| Shuffle                      | Shuffle refers to randomizing the position of sensitive data to destroy the relationship between sensitive data and achieve the purpose of desensitization. |

The commonly used desensitization strategies are replacement-based, deletion-based,
transformation-based, masking-based, encryption-based and mixed-washing-based methods [18]. The specific description of desensitization methods is shown in Table 1. Desensitization rules can be divided into two categories: recoverable and non-recoverable. The desensitization rule of recoverable class means that the desensitized data can be restored to the initial data after specific processing, such as encryption algorithm. The desensitization rule of unrecoverable class means that the desensitized data cannot be recovered, such as replacing algorithm rules. There are a large number of personal sensitive data on the CT images based on DICOM standard, such as the patient's name, gender, ID, date of birth and other personal information under the PatientTag, and the ID, type, location, time and other private information of the examination under the StudyTag.

3. Formulation of desensitization strategy

3.1. Factors to be considered in data desensitization

Data desensitization should consider how to make the desensitized data meet the application needs as much as possible under the condition of ensuring the privacy security of users, which is also a major problem faced by data desensitization. Before desensitization, the types of sensitive data, desensitization rules and application environment should be considered. The desensitization strategy should be formulated after the sensitive data type and application environment are defined. The selection of desensitization algorithm will determine the effect of desensitization. In order to develop CT image desensitization strategy to meet the above requirements, the following factors are mainly considered in this paper:

1) Availability. That is, the desensitized data can be used to meet some practical application needs. If the data cannot be used in the subsequent analysis application after processing, the data set has no use value. In the application of medical CT image, some non critical information (such as diagnosis information) can be retained after the patient's identity is anonymous, so as to meet the needs of subsequent data mining.

2) Authenticity. That is, the processed data retains some features of the original data. In order to make the desensitized data still keep the relevant characteristics of the initial data, it is usually necessary to develop a replacement method that conforms to the original data format specification to achieve the camouflage effect.

3) Timeliness. The data should be provided in a timely manner. Too long processing time may lead to a reduction in the value of data analysis and application after processing. Therefore, the desensitization algorithm should be as simple as possible under the condition of satisfying the desensitization standard.

4) Security. If the desensitized data is restored by external personnel by some means, it will cause privacy disclosure. In order to ensure the security of users' privacy, we should use substitution or generation algorithm in combination with specific scenarios.

After considering the above analysis factors, this paper develops a proper desensitization strategy, which can not only ensure the privacy of patients, but also make the desensitized data meet the application needs.

3.2. Formulation of desensitization strategy for CT images

Data attributes on CT images based on DICOM standard can be divided into three categories [19]:

1) Individual identification attributes (ID). That is, attributes that can directly indicate the identity of the individual, such as the name of the patient and the ID of the patient.

2) Quasi-identifier attribute (QI). The quasi-identification attribute itself can not directly determine the individual identity, but multiple values combined with external data may identify the individual identity. Such as gender, age, etc.

3) Sensitive attribute (ST). Specific information that describes the patient's privacy, such as diagnostic information.

According to the different data attributes of CT images, different desensitization algorithms are
adopted. Figure 1 shows the process of desensitization strategy development. The desensitization strategy developed in this paper is to replace or mask individual identity attributes, replace or delete some quasi-identity attributes, and retain the remaining non-key quasi-identity attributes and sensitive attribute information (diagnostic information). In this way, the patient's identity is kept anonymous, and sensitive information can't correspond with the patient's, so the purpose of desensitization is realized. In this way, the data on CT image still has certain validity and authenticity after desensitization, which can meet the needs of certain analysis and application.

![Desensitization procedure flow chart](image)

**Figure 1.** Desensitization procedure flow chart

4. **Overall program design**

4.1. **Development tools and data sets**

Python language is widely used now. It is a high-level programming language with dynamic semantics, interpretation and object-oriented. It plays an important role in big data and cloud computing, machine learning, artificial intelligence, image processing and other fields. Anaconda is an open source Python distribution that can manage packages and environments uniformly. It integrates the common python libraries in the field of data science and can give full play to the advantages of the Python language. The program in this paper runs under the Windows operating system. Through the Anaconda3 (Python3.7.3) interpreter, the CT image is read by the Pydicom package specially dealing with medical images in python, and the patient-related information on the CT image based on the DICOM standard is desensitized by the desensitization strategy, and finally the desensitized CT image is obtained.

The dataset used in this test is the public CT image obtained on the NBIA (National Biomedical Imaging Archive) image server.

4.2. **Program flow framework**

In this paper, the process can be divided into three parts: ① Determine the number of CT images to desensitize; ② Enter the path to the file or folder where the CT image is located; ③ Save the desensitized CT image to the determined location. The specific process of the program is as follows: the user opens the desensitization program, decides whether to desensitize a single CT image or batch desensitization, and then enters the file or folder location of the CT image; Enter the path to save the CT image after desensitization; Click the run button to start the desensitization program; Desensitization is over. The overall flow chart of this desensitization program is shown in Figure 2.
4.3. Specific function implementation

The procedure in this paper can realize the desensitization processing of single and multiple CT images based on the DICOM standard according to the formulated desensitization strategy. The specific desensitization strategy method implementation process is shown in Figure 3. Under the Anaconda integrated development environment, the CT image is read by the pydicom.read_file method of the Pydicom package that specializes in medical imaging in Python, and the corresponding methods are applied to the patient's name, gender, ID, and date of birth on the CT image. For the desensitization treatment, the corresponding desensitization strategy in this article is as follows:

1) Desensitization of gender and name was achieved by randomly generating and replacing the original values. Firstly, gender is randomly generated, and then masculine or feminine names are randomly generated according to gender. The purpose of this is to make the desensitized data more authentic and prevent the obvious contradiction between randomly generated names and gender. Then, the Patient Sex and Patient Name values under the Patient Tag are replaced with the generated values. According to the original data display format of CT image, the gender male is written in m under the Patient Sex value, and the gender female is written in F. All generated names are converted to uppercase and the Patient Name value is written.

2) Masking strategy was used for Patient ID. As the Patient ID belongs to the attribute of personal identification, it can directly correspond to the patient's identity, so it must be desensitized. In this paper, the value of the Patient ID under the Patient Tag is replaced by * instead.

3) Delete the date of birth, age and other information. The patient's date of birth, age and address are quasi-identification information. Although it is impossible to determine the patient's identity by looking at one of them alone, it is possible to infer the patient's identity by combining them with the comprehensive analysis of external data, so these informations must also be processed accordingly. In this paper, the values of PatientBirthDate, PatientSex and PatientAddress under the PatientTag are set to null to achieve desensitization effect.

After the desensitization mentioned above, the identity of the patient has been kept anonymous. In order to ensure the authenticity and availability of the data on the processed CT image, this paper decides not to deal with StudyTag, SeriesTag and other information, which can be used for follow-up.
data analysis, data cleaning and so on.

5. Result analysis
In order to show the desensitization effect of this program on CT images based on DICOM standard, firstly, select a CT image randomly in the CT image data set, and the example is shown in Figure 3. In this paper, the acquired image data are desensitized by masking and deleting strategies and desensitization strategies proposed in this paper. The final result is shown in Figure 4, where Figure 4 (a) is the result of desensitization strategy by deleting and masking, and Figure 4 (b) is the result of desensitization by the methods proposed in this paper.

![Original CT image without desensitization](image1)

**Figure 3.** Original CT image without desensitization

![CT image desensitized by masking and deleting strategy](image2)

![CT image after desensitization with the method in this paper](image3)

**Figure 4 (a).** CT image desensitized by masking and deleting strategy  
**Figure 4 (b).** CT image after desensitization with the method in this paper

From the comparison of the above results, it can be seen that the data on the picture processed by this desensitization method has more authenticity and availability, and has higher potential use value. In this paper, the identity of patients is kept anonymous after desensitization, which can achieve the purpose of desensitization, which can ensure the privacy and security of patients in the process of data mining of medical CT images, and at the same time retain the real diagnostic information of patients. Users can classify and manage CT images according to the location or time of diagnosis, which is helpful to the medical data management of radiology department.

6. Conclusion
For the security problems of patients' personal privacy information in medical big data and medical
image processing fields, and considering the use value of desensitized data, this paper formulates a strategy that can make patients’ identity anonymous, and uses Python language to design and develop a CT image desensitization method based on DICOM standard protocol. This method is implemented on the platform of Windows, which can not only process a single CT image, but also desensitize a batch of CT images. The experimental results show that this program desensitizes part of the patient information on the CT image, makes the patient identity anonymous, and the desensitization effect is good. The data on the desensitized CT image still has a certain authenticity and availability, and the desensitized data has higher use value than the previous research, and can meet the needs of subsequent data mining to a certain extent.

Acknowledgement
This paper is 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] XIAO Lian, LI Di, SUN Yang, et al. Personal Privacy Protection in the Big Data Environment of Healthcare [J]. Chinese Medical Record, 2019, 20(12): 48-50.
[2] 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.
[3] HE Yongjun, LUO Jiawei, YU Aimin. Sequencing data mining and rules analysis of multi-frame CT image data [J]. Modern Electronics Technique, 2017, 10(14): 106-108+113.
[4] 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 Usinga Fog Computing Facility With Pairing-Based Cryptography [J]. IEEE Access, 2017: 22313-22328.
[5] WANG Yang, LIU Li-bo. Research and Implementation of Desensitization System for CT Medical Images Based on DICOM [J]. Modern Computer, 2019, 72-75.
[6] 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.
[7] 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.
[8] Kanoun S, Silva Y E, Vongsalat A, et al. Automating DICOM retrieve from PACS with free and open-source software [J]. Eur J Nucl Med Mol Imaging, 2018: S298-S298.
[9] Sensakovic W F, Warden D, R, Hough M C. Troubleshooting Image Quality and Other Problems by Using the DICOM Header [J]. Radiographics, 2018, 38(3): 847-+.
[10] Bennett W, Smith K, Jarosz Q, et al. Reengineering Workflow for Curation of DICOM Datasets [J]. Digit Imaging, 2018, 31(6): 783-791.
[11] Lituitiev D S, Trivedi H, Panahia M, et al. Automatic Labeling of Special Diagnostic Mammography Views from Images and DICOM Headers [J]. Digit Imaging, 2019, 32(2): 228-233.
[12] Nooh A, Lubov J, Aoude A, et al. Differences between Manufacturers of Computed Tomography-Based Computer-Assisted Surgery Systems Do Exist: A Systematic Literature Review [J]. Glob Spine, 2017, 7(1): 83-94.
[13] El-Regaily S A, Salem M A, Aziz M H A, et al. Survey of Computer Aided Detection Systems for Lung Cancer in Computed Tomography [J]. Curr Med Imaging Rev, 2018, 14(1): 3-18.
[14] Sethi G, Saini B S. Computer aided diagnosis system for abdomen diseases in computed tomography images [J]. Biocybern Biomed Eng, 2016, 36(1): 42-55.
[15] CHEN Tian-ying, CHEN Jian-feng. Intelligent Data Masking System for Big Data Productive Environment [J]. Communications Technology, 2016, 49(07): 915-922.
[16] Price W N, Cohen I G. Privacy in the age of medical big data [J]. Nat Med, 2019, 25(1): 37-43.
[17] Mostert M, Bredenoord A L, Biesaart Mcih, et al. Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach[J].Eur J Hum Genet, 2016, 24(7): 956-960.
[18] WANG Xin, WANG Dian-gang, MU Ji-yuan, et al. Research and Implementation of Data Masking System Based on Machine Learning[J].Electric Power ICT, 2018, 16(01): 33-38.
[19] TONG Yun-Hai, TAO You-Dong, TANG Shi-Wei, et al. Identity-Reserved Anonymity in Privacy Preserving Data Publishing[J]. Journal of Software,2010, 21(04): 771-781.