SINFONIA cleaning-before-uploading guide

Preface

One of the specific challenges that the SINFONIA project pledged to address is to take into consideration risk communication and the ethics of medical applications. To achieve this, the SINFONIA proposal states that all data managed in the repository has to be stored in a secure manner and be de-identified to remove all the patient identity related information. Moreover, the de-identification has to be performed on the site of origin before sending the data to any server, in order to avoid any mitigation of personal data from the site to the remote repository.

Those statements make each SINFONIA member responsible for the files they upload to the project data repository. However, they must at the same time have in mind the aim of keeping a proper trade-off between the completeness needed to share valuable information among the project partners and the privacy requirements set by the regulation applicable. With relation to that, the project proposal remarks the commitment to take all measures to anonymize or remove personally identifiable information using state-of-the-art methods, in a manner that conforms to multi-national regulations on data protection, and particularly the General Data Protection Regulation (2016/679).

This guide intends to be a reference about a handful of tools that may help you to properly “clean” your files before uploading them to the repository. By "cleaning" we mean the process of removing and/or replacing personal information contained in your DICOM and non-DICOM files. Such a process may include also de-facing, this is, the removal or replacement of facial features in the DICOM images, to avoid 3D reconstructions being able to match personal photographs.We are rather using the verb “clean” instead of words like “anonymize” or “de-identify” because the latter ones have very specific meanings determined in both the DICOM standard and the aforementioned regulations and their full attainment is not trivial at all.

The rest of the document is organised as follows: first we review some background about the directives set by the DICOM standard regarding the protection of personal data and how a couple of research works implement that protection for DICOM-RT filesets, then some tools for cleaning DICOM and non-DICOM files are introduced, as well as some de-facing methods, and finally we provide a brief summary of the content.

Background

It is common sense to redact from your files obvious personal information like names, addresses, birth dates or identification numbers of both patients and other involved individuals (e.g.: staff, referrers, family). However, a quick review of both the GDPR and the DICOM standard will prove that common sense is not enough at all to even start a proper personal information cleaning process.

Regarding the GDPR, the Recital 35 considers as personal health information any number, symbol or particular assigned to a natural person to uniquely identify the natural person for health purposes, but also information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples; and any information on, for example, a disease, disability, disease risk, medical history, clinical treatment or the physiological or biomedical state of the data subject independent of its source, for example from a physician or other health professional, a hospital, a medical device or an in vitro diagnostic test.

With relation to the DICOM standard, the Part 15 (PS3.15) is wholly dedicated to define “Security and System Management Profiles”. This part describes in its Annex E, titled “Attribute Confidentiality Profiles”, a set of “*Profiles and Options to address the removal and replacement of Attributes within a DICOM Dataset that may potentially result in leakage of Individually Identifiable Information (III) about the patient or other individuals or organisations involved in acquisition.”* In the Table E.1-1 from this Annex there are 460 different attributes listed as affected by the different *Application Level Confidentiality Profiles* of the DICOM standard.

Let us focus on a couple of works ([Newhauser2014], [Kundu2020]) that present de-identification pipelines for DICOM-RT files. Authors of [Newhauser2014] propose the redaction of 48 attributes, whereas the process described in [Kundu2020] raises the number up to 82, with some even not being listed in the aforementioned Table E.1-1 of the standard. Furthermore, the actions to perform on some of these attributes are not limited to delete or overwrite the content with arbitrary values, since there is another table in the standard (Table E.1-1a) that defines such specific actions. Moreover, files in DICOM-RT sets are related to each other in terms of temporal and referential integrity, which makes it even more complicated to properly rewrite some of those attributes.

An authors' PDF version of [Newhauser2014] is publicly available in the academic social network ResearchGate. Table 1 of this paper includes the list of attributes and the operations to perform on it. Regarding [Kundu2020], the full table with both attributes and actions is offered by the paper's publisher as part of the preview content of the paper.

Let us note also that throughout the DICOM de-identification bibliography you may find the *Annex E* of the DICOM standard referred as *Supplement 142*, as it was born as a consequence of the original PS3.15 part of the standard did not sufficiently protect identities for patient records used in clinical trials [Newhauser2014].

DICOM cleaning tools

In this section we describe three different DICOM cleaning tools that will help partners to identify and remove the most of the personal information their files may contain in both headers and pixel data. All these tools have been installed and tested in a typical Windows desktop workstation. Furthermore, some indications are given with relation to the cleaning of non-DICOM files such as conventional images or PDF documents.

PixelMed DicomCleaner

PixelMed DicomCleaner is a free open-source tool able to clean the DICOM header of a selected set of instances, and to blackout burned-in annotations in the pixel data of the cleaned files. The development of this tool is managed by David Clunie, head of PixelMed Publishing and editor of DICOM Supplement 142 (now Annex E of PS3.15).

You can run PixelMed DicomCleaner either within your web browser or locally after downloading their files to your computer, being distributed in both Windows and MacOS versions. Basic instructions about both options are respectively given in the sections "How to install it (locally)" and "How to start it" of the tool web page.

The tool's default behaviour is to apply the *DICOM Application Level Confidentiality Profile with the Retain Longitudinal Temporal Information with Full Dates, Retain Patient Characteristics, Retain Device Identity and Retain Safe Private Options* [Clunie2016]. It does not expect the user to explicitly set which attributes must be cleaned in order to apply these or other confidentiality profiles, but to choose specific cleaning operations that modify the values of different groups of attributes. According to the tool web page, the user is provided with control over:

- Replacement values for Patient's Name, Patient's ID and Accession Number
- Modification of dates and times (e.g., Study Date), in a manner that preserves temporal relationships
- Replacement of all other identifying attributes (e.g., Referring Physician's Name, etc.)
- Removal or retention of descriptions (e.g., Study Description), which though useful may sometimes have identifying information, with separate options to remove or retain the Series Description and Protocol Name (which are generally both useful and safe)
- Removal or retention of patient characteristics (e.g., sex and weight), which are essential for PET SUV but otherwise often removed
- Removal or retention of device identifiers (e.g., serial number), which may be needed to track device performance but otherwise may be removed

- Removal or retention of institution identifiers, which may be needed to track facility performance but otherwise may be removed
- Removal or retention of clinical trial attributes, which may need to be removed for secondary re-use of clinical trial images
- Replacement of DICOM unique identifiers, which is performed consistently for a set of instances to maintain referential integrity
- Removal or retention of private attributes, except those that are known to be safe to leave and are important (e.g., SUV scale factor)
- Removal or retention of structured content, such as the content tree of DICOM Structured Report (SR) files

Apart from the tool's official web page, you may want to have a look at this short post from the blog of a private medical imaging company about how to anonymise a DICOM file. The section "Step 2: Anonymise the DICOM file" includes a set of screenshots with captions that quickly illustrate the cleaning process with Pixelmed DicomCleaner.

Moreover, some types of images contain identifying information not only in the DICOM metadata but also burned into the pixel data. Such occurrences of sensitive information need to be obscured by replacing the pixel values using an image editor. This tool embeds such an editor that can be invoked for previously cleaned files by clicking on the "Blackout" button in the main DicomCleaner screen. The editor expects the user to draw bounding boxes over the areas that need redaction, and to confirm its application over a single image or the whole cleaned fileset. You can find more information about how to proceed in the "Blackout (Redaction)" section of the DicomCleaner guide.

Once you have finished your cleaning and blackening operations, you can export the resulting DICOM files and DICOMDIR directly into a folder or into a ZIP file.

Yakami DICOM Tools Converter

The YAKAMI DICOM Tools is a set of freeware applications to handle DICOM files for research that was developed by the Department of Diagnostic Imaging and Nuclear Medicine of the Kyoto University.

It includes among other tools the DICOM Converter, which is a DICOM-to-DICOM/Image file converter that can be used to clean both headers and pixel data. You can download a ZIP containing a typical Windows installer of the toolset from here. After installing the toolset, you must look for the "DICOM Converter" application. We must warn about the quirky adaptation of the user interface from Japanese to English (misleading translations, bad support of text encoding) and the incomplete support of modern Windows features such as paths with blank spaces on folder or file names.

This is a summary of the cleaning files workflow using this tool:

1. Click on "Refer" to set a base folder to import DICOM files from
2. Use the buttons in Add panel to import either files ("File") or whole folders ("Folder")
3. Select from the "File list" panel on the left the files you want to process
4. In "*Convertion* Options" panel below, click on the "Header" tab
5. Uncheck the "specify a conversion script file" option
6. Check the "specify conversion options" option and click on "Options…"
7. Take your time to have a look at all the tabs and the rewriting/deleting actions offered
8. Once you have set the options for your cleaning process click "Save and Exit"

If you need to perform specific operations that are not directly offered by the tool, you can write your own scripts to apply them. The "Syntax" link in the "Header" tab opens a text file about how to write such scripts for custom conversions. Once you have prepared your own script, you can run it by unchecking the "specify conversion options" option, checking the "specify a conversion script file" option and selecting the script file by clicking on "refer".

This tool also offers a blackout feature in the “Image” tab of the “*Convertion* options” panel, but you must set apply a single mask at a time by explicitly setting the position top left corner and the size of the masking rectangle.

In order to apply all the operations in both the headers and the pixel data, you must go to the “Output” tab in “*Convertion* Options” and click on “Refer” to set a base output folder. You can use the “Specify output path” and the “Add extension” options to customise where and with which name and extension the cleaned files are generated. We recommend checking the “Add extension” to ensure that output files can be seamlessly opened by your default DICOM viewer. Finally, you must click on the “Convert All” button at the bottom.

RSNA MIRC Clinical Trials Processor

The Medical Imaging Resource Center (MIRC) of the Radiological Society of North America (RSNA) offers the RSNA MIRC Clinical Trials Processor (CTP), a whole suite of stand-alone image processing tools for imaging clinical trials data. It includes pre-defined implementations for key components of medical image processing pipelines, including DICOM import and anonymization. However, its full deployment may require some programming and system administration skills that we know that not every SINFONIA partner have.

Nevertheless, if you are interested in giving a try to the DICOM cleaning capabilities of CTP, you can have a look at a stand-alone version of this component. The installer is available for download in the Download Software page of the RSNA MIRC project. You need the Java Runtime Environment installed in your computer (available here) to run both the installer and the tool itself. The installer will extract some files in the folder you indicated, and the tool will open by double-clicking in “`DicomEditor.jar`”.

The workflow of this tool is similar to that of Yakami DICOM Converter: you can define your own cleaning operations over some default options already set, and also write your own cleaning scripts. After the first run of the tool, please refer to the quick reference guide provided in the “Help” tab for details about its usage.

Other tools

If you are already familiarised with any other de-identification tool (commercial, open-source, even developed by yourself or your team), feel free to use it if you consider that it performs a cleaning process robust enough to deliver files that you can confidently upload to the repository. Furthermore, if you do not feel comfortable with any of the alternatives provided in this guide, you may want to have a look at the de-identification tool survey presented in [Aryanto2015].

De-facing

The improvements in scanning and facial recognition technologies, along with the larger amount of MRI, PET and CT head images publicly available, lead to new privacy issues to consider. [Prior2008, Mazura2012, Schwarz2019] have shown that it is possible to match 3D renders of high-quality medical images with photographs. [Gao2023] argues that this matching has been done in highly-controlled, small-scale environments, where there always exists such a match and that it has not been tested yet in large sets where there might be no photograph-3D render pairings. In any case, given the continuous upgrade and refinement of technology, it is a sensible assumption to think that, eventually, Internet photos and 3D reconstructions from medical head images can be paired up, thus giving rise to new privacy breaches to be taken into account. So, even if DICOM metadata is properly de-identified, a de-facing process should also be implemented in order to minimise risk of exposing a patient’s identity.

De-facing procedures distort or conceal the patient’s facial features that most easily give away their identity in a 3D reconstruction (eyes, nose, mouth and, in the majority of cases, also ears). Due to this fact, one should be careful and think about what the DICOM files are going to be used for, as de-facing methods can mess up brain segmentation algorithms or even delete part of data meaningful for research or treatment.

In the case a de-facing procedure is needed, there are plenty of algorithms, each with its own strategy to mask the patient's face.

Pydeface

Pydeface is an open-source Python de-facing algorithm that has become popular due to its relatively easy installation and use [Gulban2019]. It comes with a predefined template atlas of facial features that is aligned to the input DICOM image using FSL's Linear Registration Tool (FLIRT) [Jenkinson2001].

This new template mask is then applied to the input image, setting to zero the corresponding facial voxels and thus effectively removing those features. However, ears are not included in the defacing mask. It does not perform any morphological operations to the transformed facial mask.

Afni_refacer

Afni_refacer is a de-facing tool part of AFNI (Analysis of Functional NeuroImages) [Cox1996], a software suite of C, Python, R programs as well as shell scripts developed for the analysis and display of MRI images.

It has two operation modes. In the first one, it follows the same steps as Pydeface, with a pre-defined mask aligned using AFNI 3dAllineate and a MNI template, that is later used to remove facial voxels, including ears. The other operation mode replaces the facial features with artificial values. This latter mode has two possible outputs: a refacing of face and ears or also including a replacement of the skull.

This software is freely available for research purposes. It runs on virtually any Unix system with X11 and Motif displays. It can be found as open-source code or precompiled binary packages, provided for MacOs and Linux systems (Fedora, CentOS/Red Hat and Ubuntu, which allows users to use it in Windows by means of its Windows Subsystem for Linux).

AnonyMI

AnonyMI is an open-source tool for defacing MRI images that follows three main steps [Mikulan2021]. First, it obtains a 3D reconstruction of the subject's skin and skull through a watershed algorithm [Dale1999, Ségonne2004]. Then, it performs a nonlinear registration between a template (from the IXI dataset [Avants2008, Avants2018]) and the input image to locate the face and ears, the control points. This step allows for some personalisation, in case the subject's identification requires larger areas to be anonymised (e.g. due to scars). Finally, a subject-specific mask is made using the skin and skull surfaces from the first step and the control points. In the places of intersection between the surfaces and the control points, the mask is filled with random numbers that follow the same intensity distribution as the voxels between the 3D skin and skull reconstructions. Outside the skin surface values are set to zero. The resulting volume of these three steps does not substantially modify the geometrical properties of the input image. This program provides a graphical interface, but it can be also used as a command line script.

Mri_reface

Mri_reface is a de-facing method that, rather than removing facial and ears voxels, replaces them with those of an average face. It also replaces some regions of air, as they may introduce some identifiable features due to wraparound artefacts. This way the resulting output better resembles a natural image and reduces on downstream brain measurement software [Schwarz2021, Schwarz2022]. It uses an average face template that is registered with the input image using ANTs [Avants2008], linearly on face and ears, nonlinearly on the other regions. The template, after application of a deformation field, only preserves the original brain data. The blending of the new facial features and ears is done by matching the image intensities of the template and input images by means of a combination of global and piecewise intensity matching and bias correction for smooth local intensity normalisation between the images [Vemuri2015].

At the moment, the mri_reface software is compiled matlab only and requires Linux with the (free) matlab runtime installed. It is freely available only for non-commercial research use.

Other tools and models

Many other de-facing models are already publicly available: MiDeFace, mri_deface, Quickshear, FSL_deface, Face_Masking, BrainVoyager… Models are dataset dependent, but the four ones previously detailed seem to give better results, according to the comparisons made in [Theyers2021, Gao2023].

What about non-DICOM files?

We are aware that in some cases your DICOM filesets may be linked to non-DICOM files, which also should be cleaned before uploading them to the SINFONIA repository. Personal information contained in typical image files (TIFF, JPG, PNG…) can be obscured using an image editor like the well-known GIMP or even Microsoft Paint, whereas PDFs files (e.g.: medical reports) can be redacted using the Adobe Acrobat commercial suite if you have access to a licence. Unfortunately, we are not aware of any suitable open-source or freeware alternative for it at the current moment.

Summary

In this document we remind the significance that a proper management of personal information has within the SINFONIA project and the responsibility that partners have with relation to that as they are commanded to remove all the patient identity related information from any file they want to upload to the repository. A quick review about what is considered as identity related information on both the EU regulation (GDPR) and the technical standards (DICOM) proves that common sense is not enough to fulfil this requirement. Because of that, some cleaning tools are introduced in order to help the partners to clean the personal information that their files may contain. A brief description and some usage guidance is provided for three DICOM cleaning tools, along with some comments about how to deal with non-DICOM files.

Research papers referenced

[Newhauser2014]

Newhauser, Wayne & Jones, Timothy & Swerdloff, Stuart & Newhauser, Warren & Cilia, Mark & Carver, Robert & Halloran, Andy & Zhang, Rui. (2014). **Anonymization of DICOM Electronic Medical Records for Radiation Therapy**. Computers in Biology and Medicine. 53. https://doi.org/10.1016/j.compbiomed.2014.07.010.

[Kundu2020]

Kundu, Surajit & Chakraborty, Santam & Chatterjee, S. & Das, Syamantak & Basu, Rimpa & Mukhopadhyay, Jayanta & Das, Parthapratim & Mallick, Indranil & Moses, Arunsingh & Bhattacharyyaa, Tapesh & Ray, Soumendranath. (2020). **De-Identification of Radiomics Data Retaining Longitudinal Temporal Information**. Journal of Medical Systems. 44. https://doi.org/10.1007/s10916-020-01563-0.

[Clunie2016]

Clune, David. **Letter to the editor: "Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy"**. European Radiology Opinions, April 20th, 2016. https://www.european-radiology.org/opinions/clunie-2016/.

[Aryanto2015]

Aryanto, K.Y.E., Oudkerk, M. & van Ooijen, P.M.A. **Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy**. European Radiology 25, 3685–3695 (2015). https://doi.org/10.1007/s00330-015-3794-0.

[Prior2008]
F. W. Prior, B. Brunsden, C. Hildebolt, et al. **Facial recognition from volume-rendered magnetic resonance imaging data.** IEEE Transactions on Information Technology in Biomedicine. 13(1), 5–9 (2008). https://doi.org/10.1109/TITB.2008.2003335.

[Mazura2012]
J. C. Mazura, K. Juluru, J. J. Chen, et al., **Facial recognition software success rates for the identification of 3D surface reconstructed facial images: implications for patient privacy and security**. Journal of digital imaging 25, 347–351 (2012). https://doi.org/10.1007/s10278-011-9429-3.

[Schwarz2019]
C. G. Schwarz, W. K. Kremers, T. M. Therneau, et al. **Identification of Anonymous MRI Research Participants with Face-Recognition Software**. New England Journal of Medicine 381, 1684–1686 (2019). https://doi.org/10.1056/NEJMc1908881.

[Gao2023]
Gao C, Landman BA, Prince JL, Carass A. **Reproducibility evaluation of the effects of MRI defacing on brain segmentation**. J Med Imaging (Bellingham). 2023 Nov;10(6):064001. https://doi.org/10.1117/1.JMI.10.6.064001.

[Gulban2019]
Omer Faruk Gulban, Dylan Nielson, Russ Poldrack, John Lee, Chris Gorgolewski, Vanessasaurus, & Satrajit Ghosh. (2019). **poldracklab/pydeface: v2.0.0 (v2.0.0)**. Zenodo. https://doi.org/10.5281/zenodo.3524401.

[Jenkinson2001].
M. Jenkinson and S. Smith. **A global optimisation method for robust affine registration of brain images**. Medical Image Analysis 5(2), 143–156 (2001). https://doi.org/10.1016/s1361-8415(01)00036-6.

[Cox1996]
Cox RW. **AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages**. *Comput Biomed Res.* (1996) 29:162–173. https://doi.org/10.1006/cbmr.1996.0014.

[Mikulan2021]
Mikulan, E., Russo, S., Zauli, F.M., d'Orio, P., Parmigiani, S., Favaro, J., Knight, W., Squarza, S., Perri, P., Cardinale, F., Avanzini, P., Pigorini, A., 2021. A comparative study between state-of-the-art MRI deidentification and AnonyMI, a new method combining re-identification risk reduction and geometrical preservation. Human Brain Mapping 42, 5523–5534. https://doi.org/10.1002/hbm.25639

[Dale1999]
Dale AM, Fischl B, Sereno MI. **Cortical surface-based analysis. I. Segmentation and surface reconstruction**. Neuroimage. 1999 Feb;9(2):179-94. https://doi.org/10.1006/nimg.1998.0395.

[Ségonne2004]
Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B. **A hybrid approach to the skull stripping problem in MRI**. Neuroimage. 2004 Jul;22(3):1060-75. https://doi.org/10.1016/j.neuroimage.2004.03.032.

[Avants2008]

B. B. Avants, C. L. Epstein, M. Grossman, et al. **Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain.** Medical Image Analysis 12(1), 26–41 (2008). https://doi.org/10.1016/j.media.2007.06.004.

[Avants2018]
B. Avants and N. Tustison. **ANTs/ANTsR Brain Templates**. figshare. Dataset. (2018).

[Schwarz2021]
C. G. Schwarz, W. K. Kremers, H. J. Wiste, et al. **Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives**. NeuroImage 231, 117845 (2021). https://doi.org/10.1016/j.neuroimage.2021.117845

[Schwarz2022]
Schwarz CG, Kremers WK, Lowe VJ, Savvides M, Gunter JL, Senjem ML, Vemuri P, Kantarci K, Knopman DS, Petersen RC, Jack CR Jr. **Alzheimer's Disease Neuroimaging Initiative. Face recognition from research brain PET: An unexpected PET problem**. Neuroimage. 2022 Sep; 258:119357. https://doi.org/10.1016/j.neuroimage.2022.119357.

[Vemuri2015]
Vemuri P, Senjem ML, Gunter JL, Lundt ES, Tosakulwong N, Weigand SD, Borowski BJ, Bernstein MA, Zuk SM, Lowe VJ, Knopman DS, Petersen RC, Fox NC, Thompson PM, Weiner MW, Jack CR Jr. **Alzheimer's Disease Neuroimaging Initiative. Accelerated vs. unaccelerated serial MRI based TBM-SyN measurements for clinical trials in Alzheimer's disease**. Neuroimage. 2015 Jun; 113:61-9. https://doi.org/10.1016/j.neuroimage.2015.03.026.

[Theyers2021]
Theyers AE, Zamyadi M, O'Reilly M, Bartha R, Symons S, MacQueen GM, Hassel S, Lerch JP, Anagnostou E, Lam RW, Frey BN, Milev R, Müller DJ, Kennedy SH, Scott CJM, Strother SC, Arnott SR. **Multisite Comparison of MRI Defacing Software Across Multiple Cohorts**. Front Psychiatry. 2021 Feb 24; 12:617997. https://doi.org/10.3389/fpsyt.2021.617997.