Data Retrieval from PACS for BreastScreen Norway

`https://github.com/mikevoets/dicom_anonymizer`

Mike Voets, Lars Ailo Bongo
Department of Computer Science
UiT - The Arctic University of Norway
mvo010@post.uit.no, lars.ailo.bongo@uit.no

1 Introduction

We implemented an extraction and anonymization script for mammograms in Norwegian hospitals for BreastScreen Norway. From a practical point of view, it is easiest to execute the script on the hospital ICT infrastructure and to let the script consume a list of identifiers from the Cancer Registry of Norway [1] (from here referred to as the Cancer Registry). We use the script in a pilot project to extract anonymized data from the University Hospital of North Norway (UNN). We collaborate with Helse Nord IKT. Helse Nord IKT operates the ICT infrastructure of the North Norwegian health region. The project is part of a larger ICT research project at UiT - The Arctic University of Norway aiming to develop the needed infrastructure for integrated analysis of medical data.

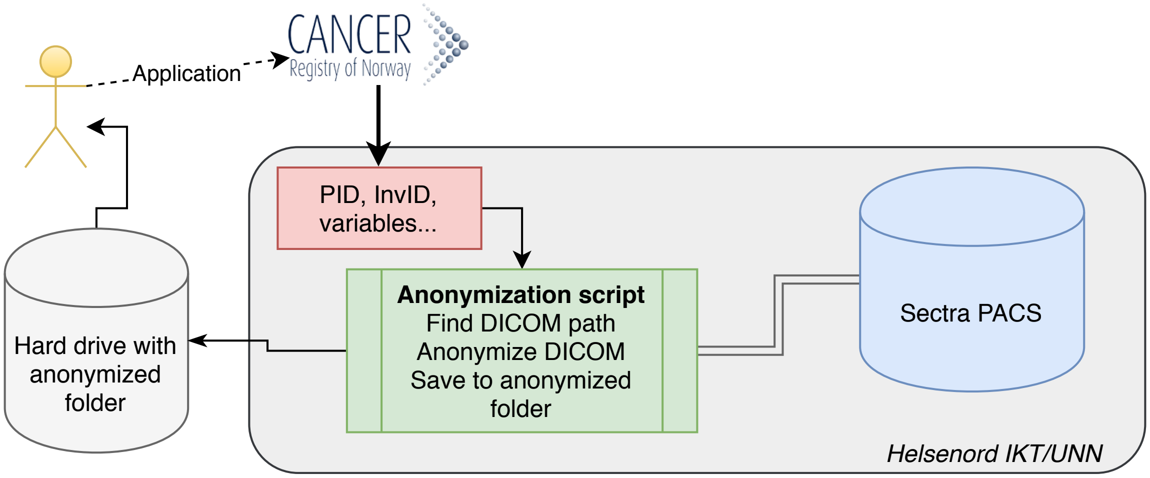

Figure 1: Overview of the mammograms anonymization process.

The mammography data set includes mammograms taken in the period from 2012 until 2018, and excludes people who have opted-out of their images being used for research. Estimations are that this part of the data set contains 280 000 images from 70 000 screenings. The mammograms reside in a *Sectra* PACS system. This system does not provide an API for extraction or anonymization of files, making the objective of our script to extract mammograms from the part of the file system used by the PACS. The mammograms are *DICOM* files. They contain personal meta-data, requiring the script to anonymize these files since there is no explicit need and allowance to use person-related data to develop a deep learning algorithm. The Cancer Registry has provided a variable specification for additional meta-data associated with each mammogram that are not part of the DICOM file, but are essential to develop a deep learning algorithm. Because these meta-data contain personal data as well, the script also anonymizes these meta-data. The script further assures that the anonymized data cannot be linked back to the original personal data.

Retrieving the variable specification and confirmation on being able to extract mammograms is an ongoing part of a lengthy application process. We do not gain access to the PACS system to extract mammograms directly, instead we wrote a prepared script that Helse Nord IKT will verify and finish before executing it in the *Sectra* PACS environment at UNN (see Figure 1). Ultimately, the script will be used in the larger ICT project to extract mammograms from all screening points in Norway.

1.1 Breast Cancer Screening

Figure 2 shows the process of breast cancer screening until periodical import into the Cancer Registry. Every business day in the North Norwegian region, about 65 people attend mammography screening at UNN, and an additional 65 people are screened remotely. The Cancer Registry imports data from UNN periodically, typically once a year. The Cancer Registry imported data from 435 000 people who were screened for breast cancer in the screening rounds of the year 2014-2015.

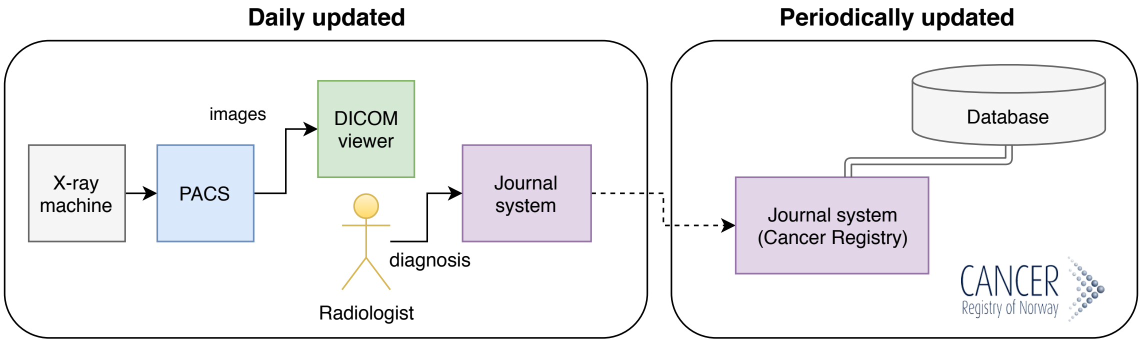

Figure 2: Visualization of breast cancer screening and periodic retrieval by the Cancer Registry.

2 Implementation

The anonymization script was written in Python 2.7, and can be used from the command line. The script accepts several parameters. First, the root directory path where the DICOM files containing digital mammograms and meta-data reside. Second, the path to the *csv* file with meta-data from the Cancer Registry. Third, the directory path where anonymized DICOM files should be placed to. Fourth, the path where the anonymized variables (cleaned file) from the Cancer Registry should be placed to.

2.1 Anonymization of DICOM files

DICOM, *Digital Imaging and Communications in Medicine*, is a standard digital file format for medical images [2, 3]. These files contain raw image data and other meta-data related to the image. This meta-data usually consists of personal information, information about the owner of the image, and information about for what purpose, when, and with what equipment the image was taken. To anonymize the personal data in the DICOM files, the script uses the *dicom-anon* Python tool [4]. Dicom-anon has been implemented by the Children's Hospital of Philadelphia (CHOP) [5]. See Table 2 in Section 2.3 for an overview on how DICOM files in our script are anonymized. Dicom-anon attempts to be compliant with the Basic Application Confidentiality Profile as specified in DICOM 3.15 Annex E document [6]. These specifications define what values in the meta-data should be anonymized based on their modality. Modality represents the DICOM file type. For mammography, the modality is *mg*. Dicom-anon further removes all attributes from the DICOM file that are not specified in Annex E. The tool creates a *sqlite3* database file with a table containing the original and cleaned version of every attribute. This file will be removed after running our anonymization script.

2.2 Anonymization of Cancer Registry Meta-Data

The scripts accepts a *csv* file with a list of variables from the Cancer Registry. See Table 1 in Section 2.3 for an overview on how the values from this file are anonymized and placed to a cleaned file. The first two variables per line represent the personal identifier (PID), and the invitation or screening identifier (InvID), respectively. PID can be linked to many InvIDs. The third value represents the screening date. The seventh value represents the diagnosis date. Other examples of values in the file are annotations for ground truth, and where the image was taken. These other values do not need to be modified, because they cannot be linked with a person's identity. The PID and InvID are not included in the anonymized meta-data file, but are used to identify the association between the person and screenings later. The screening date and diagnosis date are originally formatted as *15.mmm.yyyy.* The anonymization script re-formats the screening date to *m-yyyy*, and the diagnosis date is converted to the amount of days after the screening date. The screening date is used for the new directory structure of the anonymized DICOM files per person.

2.3 Anonymity Assurance Test

The script provides a test to assure that the resulting anonymized mammograms and meta-data cannot be linked back to their original personal information. To facilitate this test, we provided a couple of example DICOM files with fake personal information, together with an example *csv* file that represents the meta-data file from the Cancer Registry. The folder structure for the example DICOM files before running the test is shown in Figure 3. The test can be run by specifying the *-t* flag when executing the script. After running the script in test mode, the *tests* folder is modified as shown in Figure 4. The result of anonymizing a file representing the file from the Cancer Registry is shown in Table 1, and the result of anonymizing DICOM files is shown in Table 2.

3 Discussion

To implement the anonymization script, we had to make some assumptions. First, we assume that one PID can be associated with many InvIDs, because a person may be screened for breast cancer multiple times. Second, we assume that the variables in the *csv* file from the Cancer Registry are delimited by white spaces. The delimiter can however be changed in the script. Third, we assume that the variables are delimited in the same order as the variables in the variable specification received from the Cancer Registry: i.e. the first two values in the line should be *PID* and *InvID*, the third value *O2_Bildetakningsdato*, and the seventh value *Diagnosedato*.

3.1 Limitations

DICOM files that are explicitly marked as containing burnt-in data along with files that have a series description of *Patient Protocol*, will be copied to a *quarantine* folder, and cannot be anonymized by our script.

We did not know any details regarding the internal folder structure of the PACS system. Because of this, we have not implemented the method for retrieving the internal path to a specific DICOM file given a PID and InvID. Before executing this script in the *Sectra* PACS environment at UNN, Helse Nord IKT is required to verify the script and implement the method to find the internal DICOM paths for all screened people.

Original values *(in variables.csv)*

PID	**InvID**	**O2_Bildetakingsdato**	**...**	**Diagnosedato**
		Screening date		***Diagnosis date***
Example_Patient	Screening_1	15.Jan.2016	...	15.Feb.2016
Example_Patient	Screening_2	15.Dec.2017	...	15.Jan.2018

Anonymized values *(in cleaned_variables.csv)*

Anonymized PID	**Screening date**	**...**	**Diagnosis**
			Days offset
e3b23d103c4342...	12-2017	...	31
e3b23d103c4342...	1-2016	...	31

Table 1: Overview of anonymization of values of the Cancer Registry. The personal identifier *PID* is anonymized by assigning a pseudo-randomized UUID. The screening identifier *InvID* is removed. Instead, the screening date is used and formatted to *m-yyyy*. Note that this corresponds to the anonymized folder structure in Figure 4. The diagnosis date is converted to the offset in days relative from the screening date. All other variables in the file are unchanged (not shown in this table).

DICOM Attribute Name	**Value in *Screening_1/EE06C00F.dcm***	**Value in *1-2016/1.dcm***
Specific Character Set	’ISO_IR 100’	<removed attribute>
Image Type	[’ORIGINAL’, ’PRIMARY’, ”]	[’ORIGINAL’, ’PRIMARY’, ”]
Study Date	’20140408’	’19010101’
Content Date	’20140408’	’19010101’
Study Time	’104011’	’000000.00’
Content Time	’104117.000000’	’000000.00’
Accession Number	’R9BF8PC1GE’	’Accession Number 1’
Patient ID	’R9BF8PC1GE’	’Patient ID 1’
[Examination Number]	’E9BF8PC1GE’	<removed attribute>
Patient Name	’Anonymous Female 1959’	”Patient’s Name 1”
Patient’s Birth Date	’19591221’	’19010101’
Patient’s Sex	’F’	’CLEANED’
Patient’s Birth Name	’anonymous’	<removed attribute>
Patient’s Age	’054Y’	<removed attribute>
Patient’s Mother’s Birth Name	’anonymous’	<removed attribute>
Medical Alerts	’anonymous’	<removed attribute>
Allergies	’anonymous’	<removed attribute>
	...	
Study ID	’E9BF8PC1GE’	’CLEANED’
Patient Identity Removed	<non-existent>	’YES’
	...	
KVP	’30’	’30’
Distance Source to Detector	’660’	’660’
Distance Source to Patient	’660’	’660’
Estimated Radiographic Magnification	’1’	’1’
Field of View Dimension(s)	[’306’, ’239’]	[’306’, ’239’]
Exposure Time	’785’	’785’
X-Ray Tube Current	’62’	’62’
Exposure	’49’	’49’
Expore in uAs	’48800’	’48800’
	...	
Pixel Data	Array of 14660856 bytes	Array of 14660856 bytes

Table 2: Overview of DICOM meta-data anonymization. Files are anonymized by the *dicom-anon* tool [4]. For this example we used the attribute values of *EE06C00F.dcm* and its anonymized variant *1.dcm.* All person-related meta-data are anonymized by assigning a sequence. All dates are reset to 1901-01-01. *Patient’s Sex* and *Study ID* attributes are cleaned. Optional or unrecognized attributes are removed. A new attribute *Patient Identity Removed* is added to the anonymized DICOM file. The actual image represented by *Pixel Data* stays unchanged. We do not show DICOM tag and VR in this table, as they do not provide additional relevant information about the anonymization procedure.

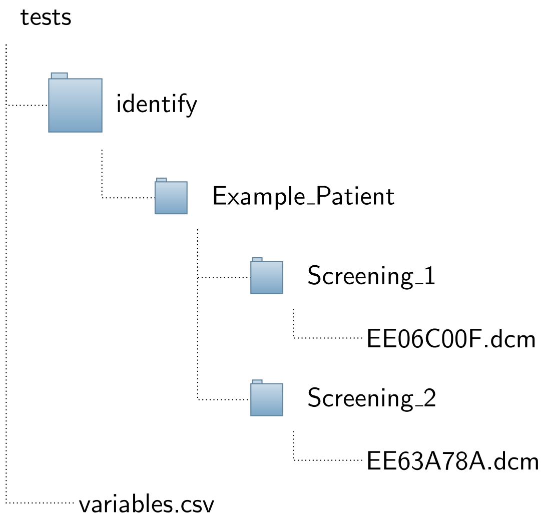

Figure 3: Before executing test. The personal folders reside in the *identify* folder (may be PID), each consisting of one or more screening (may be InvID) folders consisting of DICOM files. The *variables.csv* file represents a possible *csv* file with example variables from the Cancer Registry.

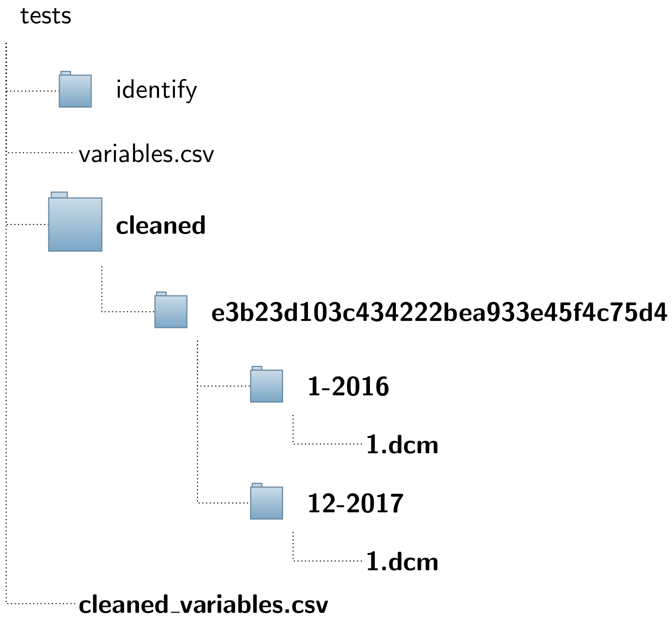

Figure 4: After having executed test. The original *identify* folder structure still exists with its original content, but a new folder *cleaned* has been created. This folder consists of the anonymized data. It consists of folders named with pseudo-randomly generated UUIDs, representing people. Each folder consists of one or more screening folders named with the screening's date formatted by *m-yyyy*, found among the Cancer Registry variables, with one or more renamed anonymized DICOM files for the corresponding screening. The cleaned variables from the Cancer Registry are placed in *cleaned_variables.csv*.

3.2 Related Work

The Digital Mammography DREAM challenge [7] was a machine learning competition held in 2016, as an attempt to find a machine learning algorithm that improves the predictive accuracy of digital mammography for the early detection of breast cancer, with its main focus on reducing the recall rate for breast cancer screening. It provided a data set consisting of 640 000 de-identified mammograms from 86 000 people with corresponding personal characteristics and outcome measures. This shows that large amounts of annotated data are needed to develop and evaluate a deep learning algorithm, and confirms that de-identification or anonymization of the digital mammograms is necessary.

3.3 Conclusion

When the script has been successfully executed, the folder with the anonymized data can be transferred from the sensitive data environment, and the resulting anonymized data can then be used to develop a deep learning algorithm.

References

[1] "The Cancer Registry of Norway." [Online]. Available: https://www.kreftregisteret.no/en/General/About-the-Cancer-Registry/About-the-organization/

[2] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *European Radiology*, vol. 12, no. 4, pp. 920–927, 2002. doi: 10.1007/s003300101100

[3] D. Peck, "Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide," *Journal of Nuclear Medicine*, vol. 50, no. 8, pp. 1384–1384, 2009. doi: 10.2967/jnumed.109.064592

[4] CHOP, "Dicom-anon: Python DICOM Anonymizer," 2016. [Online]. Available: https://github.com/chop-dbhi/dicom-anon

[5] ——, "Children's Hospital of Philadelphia." [Online]. Available: http://www.chop.edu/about-us

[6] National Electrical Manufacturers Association, "Digital Imaging and Communications in Medicine (DICOM) Part 15: Security and System Management Profiles," pp. 85–92, 2011. [Online]. Available: ftp://medical.nema.org/medical/dicom/2011/11_15pu.pdf

[7] Sage Bionetworks, "The Digital Mammography DREAM Challenge," 2016. [Online]. Available: https://www.synapse.org/#!Synapse:syn4224222/wiki/401745