NIH NATIONAL CANCER INSTITUTE

Virtual Workshop on Medical Image De-Identification (MIDI)

May 22-23, 2023
10 am – 2 pm EDT

Synthetic Data for De-Identification Testing

The MIDI Datasets

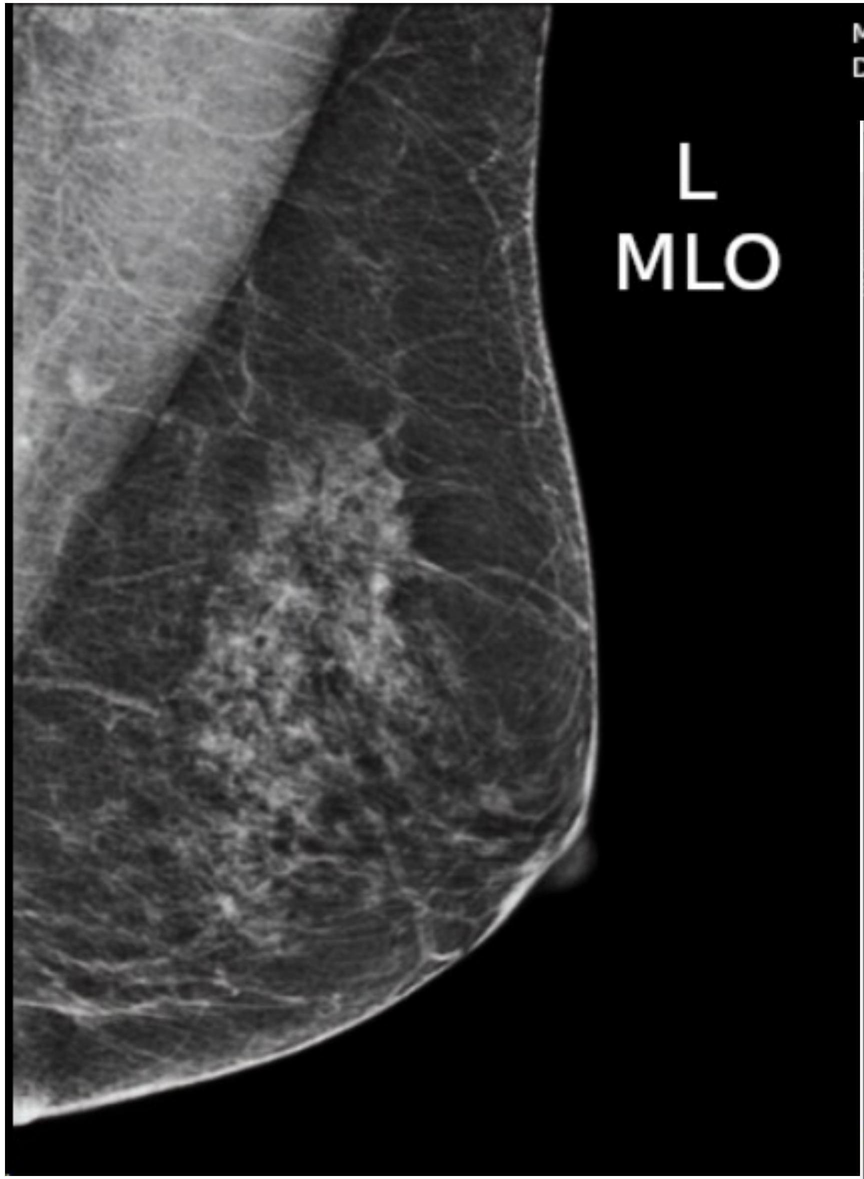
Fred Prior, PhD
Professor and Chair
Department of Biomedical Informatics
Professor of Radiology
University of Arkansas for Medical Sciences

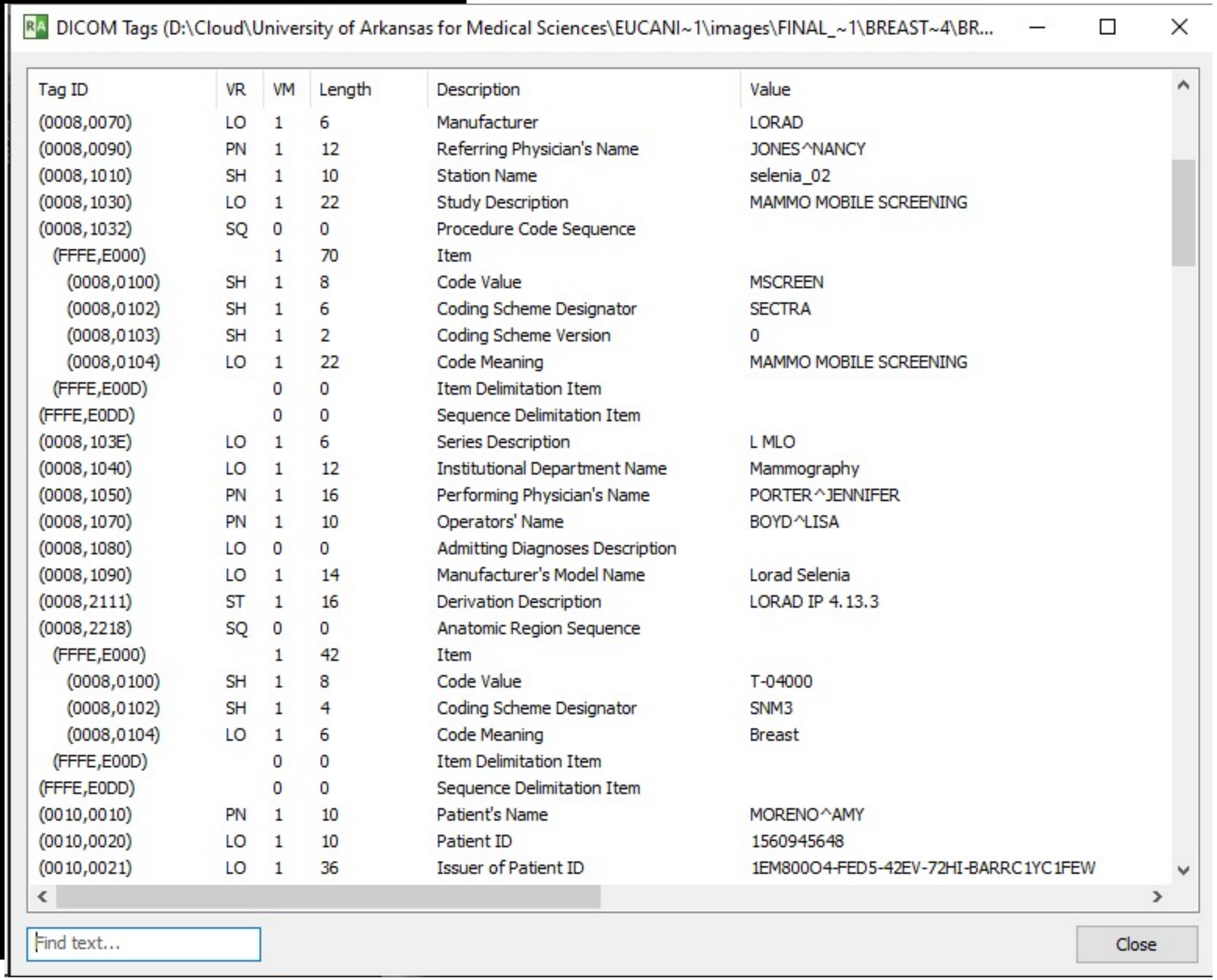
UAMS®
College of Medicine

How to Validate De-identification Tools and NOT Violate HIPAA

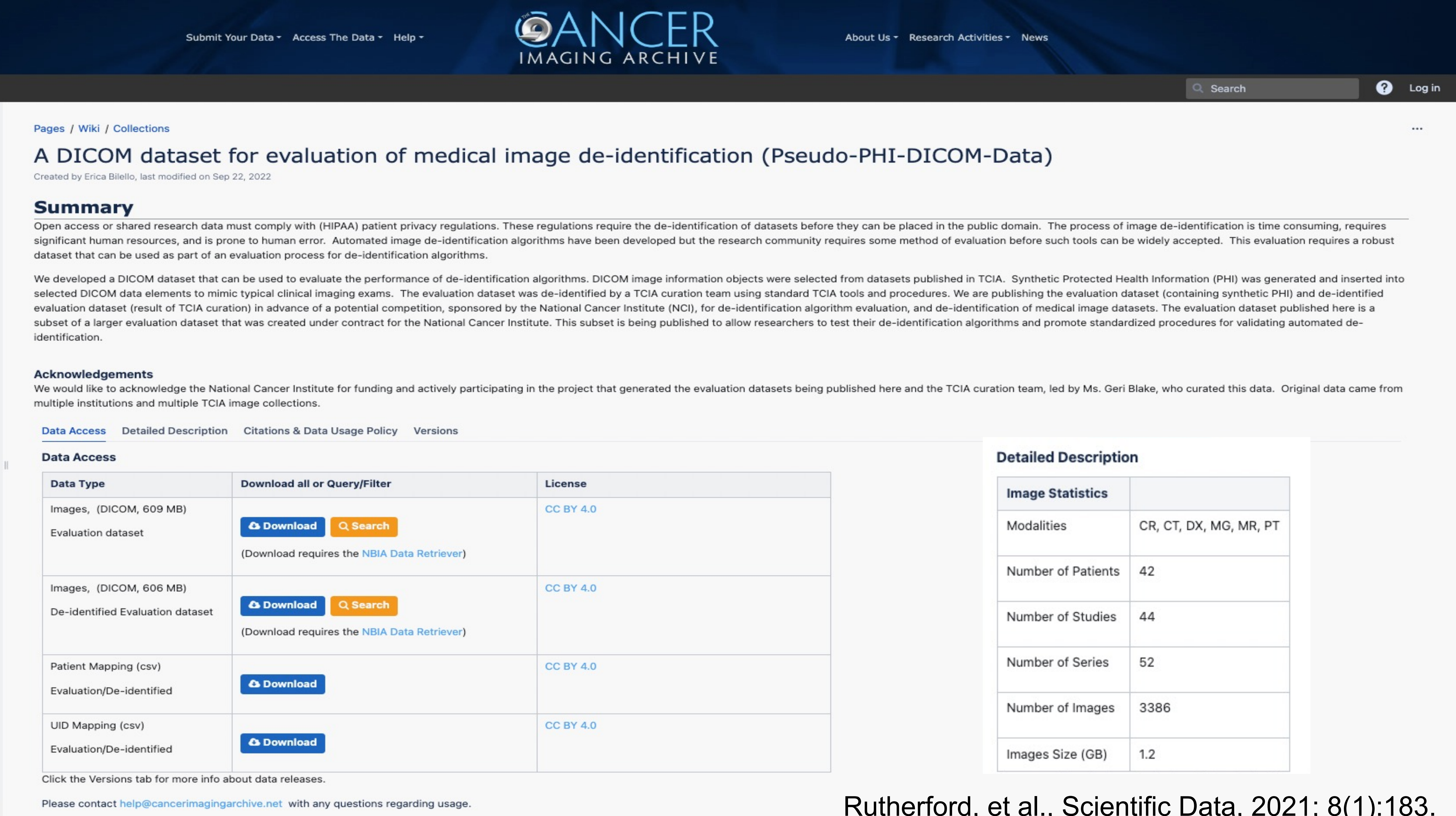
- Image anonymization algorithms and pipelines must be validated before they are deployed to process data that will be publicly shared.
- Validation requires a robust dataset (or datasets) that can be used in the assessment of de-identification algorithms.
- Synthetic datasets can be constructed to test normal and edge cases and cover all DICOM defined data object types, and images in non-DICOM formats

Synthetic DICOM Image Object: Header + Pixels




DICOM Tags (D:\Cloud\University of Arkansas for Medical Sciences\EUCANI~1\images\FINAL_~1\BREAST~4\BR...
Find text...
Close

Tag ID	VR	VM	Length	Description	Value
(0008,0070)	LO	1	6	Manufacturer	LORAD
(0008,0090)	PN	1	12	Referring Physician's Name	JONES^NANCY
(0008,1010)	SH	1	10	Station Name	selenia_02
(0008,1030)	LO	1	22	Study Description	MAMMO MOBILE SCREENING
(0008,1032)	SQ	0	0	Procedure Code Sequence	
(FFFE,E000)		1	70	Item	
(0008,0100)	SH	1	8	Code Value	MSCREEN
(0008,0102)	SH	1	6	Coding Scheme Designator	SECTRA
(0008,0103)	SH	1	2	Coding Scheme Version	0
(0008,0104)	LO	1	22	Code Meaning	MAMMO MOBILE SCREENING
(FFFE,E00D)		0	0	Item Delimitation Item	
(FFFE,E0DD)		0	0	Sequence Delimitation Item	
(0008,103E)	LO	1	6	Series Description	L MLO
(0008,1040)	LO	1	12	Institutional Department Name	Mammography
(0008,1050)	PN	1	16	Performing Physician's Name	PORTER^JENNIFER
(0008,1070)	PN	1	10	Operators' Name	BOYD^LISA
(0008,1080)	LO	0	0	Admitting Diagnoses Description	
(0008,1090)	LO	1	14	Manufacturer's Model Name	Lorad Selenia
(0008,2111)	ST	1	16	Derivation Description	LORAD IP 4.13.3
(0008,2218)	SQ	0	0	Anatomic Region Sequence	
(FFFE,E000)		1	42	Item	
(0008,0100)	SH	1	8	Code Value	T-04000
(0008,0102)	SH	1	4	Coding Scheme Designator	SNM3
(0008,0104)	LO	1	6	Code Meaning	Breast
(FFFE,E00D)		0	0	Item Delimitation Item	
(FFFE,E0DD)		0	0	Sequence Delimitation Item	
(0010,0010)	PN	1	10	Patient's Name	MORENO^AMY
(0010,0020)	LO	1	10	Patient ID	1560945648
(0010,0021)	LO	1	36	Issuer of Patient ID	1EM800O4-FED5-42EV-72HI-BARRC1YC1FEW



A DICOM dataset for evaluation of medical image de-identification (Pseudo-PHI-DICOM-Data)

Created by Erica Bilello, last modified on Sep 22, 2022

Summary

Open access or shared research data must comply with (HIPAA) patient privacy regulations. These regulations require the de-identification of datasets before they can be placed in the public domain. The process of image de-identification is time consuming, requires significant human resources, and is prone to human error. Automated image de-identification algorithms have been developed but the research community requires some method of evaluation before such tools can be widely accepted. This evaluation requires a robust dataset that can be used as part of an evaluation process for de-identification algorithms.

We developed a DICOM dataset that can be used to evaluate the performance of de-identification algorithms. DICOM image information objects were selected from datasets published in TCIA. Synthetic Protected Health Information (PHI) was generated and inserted into selected DICOM data elements to mimic typical clinical imaging exams. The evaluation dataset was de-identified by a TCIA curation team using standard TCIA tools and procedures. We are publishing the evaluation dataset (containing synthetic PHI) and de-identified evaluation dataset (result of TCIA curation) in advance of a potential competition, sponsored by the National Cancer Institute (NCI), for de-identification algorithm evaluation, and de-identification of medical image datasets. The evaluation dataset published here is a subset of a larger evaluation dataset that was created under contract for the National Cancer Institute. This subset is being published to allow researchers to test their de-identification algorithms and promote standardized procedures for validating automated de-identification.

Acknowledgements

We would like to acknowledge the National Cancer Institute for funding and actively participating in the project that generated the evaluation datasets being published here and the TCIA curation team, led by Ms. Geri Blake, who curated this data. Original data came from multiple institutions and multiple TCIA image collections.

Data Access

Data Type	Download all or Query/Filter	License
Images, (DICOM, 609 MB) Evaluation dataset	Download Search (Download requires the NBIA Data Retriever)	CC BY 4.0
Images, (DICOM, 606 MB) De-identified Evaluation dataset	Download Search (Download requires the NBIA Data Retriever)	CC BY 4.0
Patient Mapping (csv) Evaluation/De-identified	Download	CC BY 4.0
UID Mapping (csv) Evaluation/De-identified	Download	CC BY 4.0

Click the Versions tab for more info about data releases.

Please contact help@cancerimagingarchive.net with any questions regarding usage.

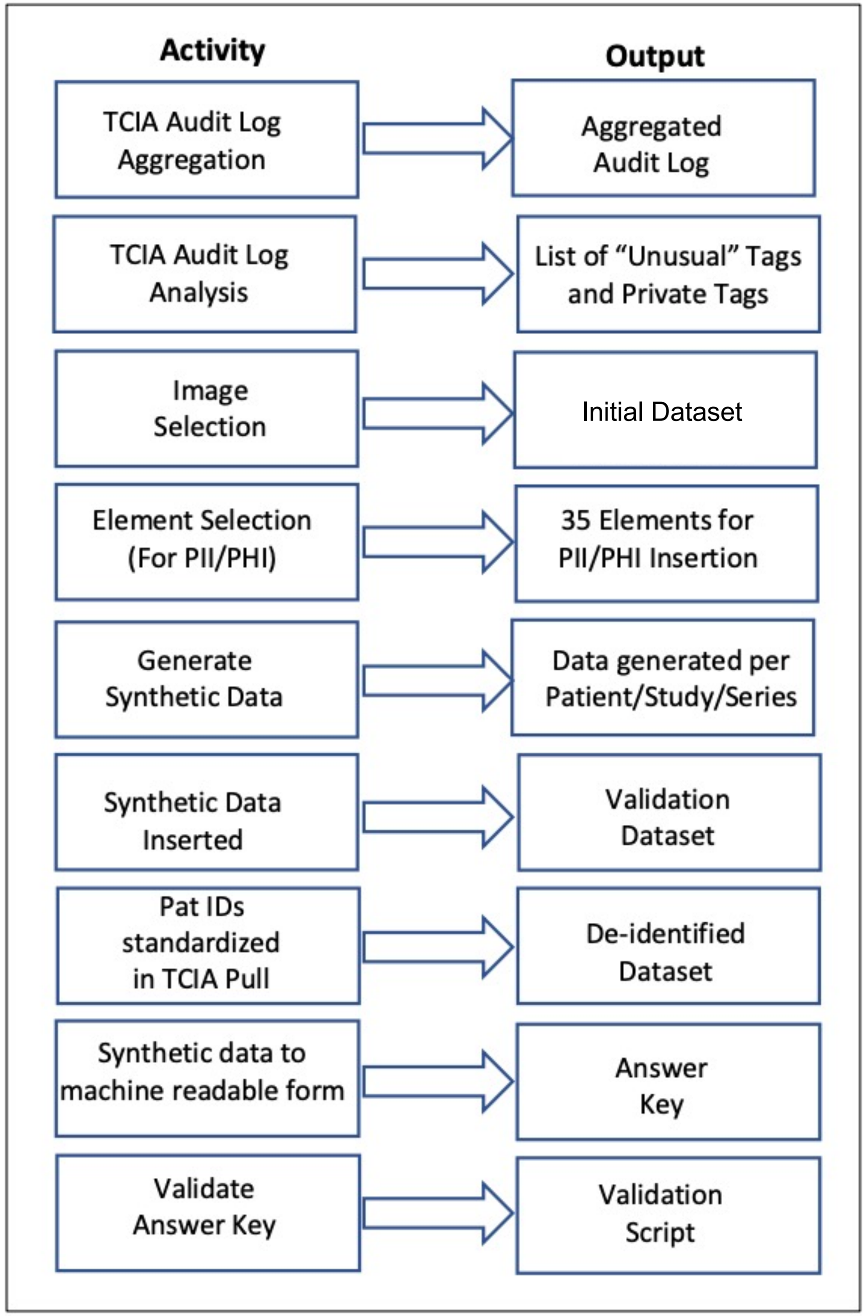
Detailed Description

Image Statistics	
Modalities	CR, CT, DX, MG, MR, PT
Number of Patients	42
Number of Studies	44
Number of Series	52
Number of Images	3386
Images Size (GB)	1.2

Rutherford, et al., Scientific Data, 2021; 8(1):183.

De-Identification Validation Dataset

- We’ve developed multiple DICOM datasets that can be used to evaluate the performance of de-identification algorithms.
- DICOM objects were selected from datasets published in TCIA.
- Synthetic PHI was generated and inserted into selected DICOM attributes to mimic typical clinical imaging exams.
- The DICOM Standard and TCIA curation audit logs guided the insertion of synthetic PHI into standard and non-standard DICOM data elements.
- TCIA curation tools and procedures were used to de-identify the synthetic data.
- An “Answer Key” was created to identify what elements should be modified during **a curation process equivalent to that used by TCIA**
- A Python evaluation script was created to compare the answer key to a de-identified dataset.



Activity	Output
TCIA Audit Log Aggregation	Aggregated Audit Log
TCIA Audit Log Analysis	List of “Unusual” Tags and Private Tags
Image Selection	Initial Dataset
Element Selection (For PII/PHI)	35 Elements for PII/PHI Insertion
Generate Synthetic Data	Data generated per Patient/Study/Series
Synthetic Data Inserted	Validation Dataset
Pat IDs standardized in TCIA Pull	De-identified Dataset
Synthetic data to machine readable form	Answer Key
Validate Answer Key	Validation Script

Validation Datasets

- A total of 172,887 images and other DICOM Object Instances representing 1,448 studies for 1,517 synthetic patients
- 28 equipment vendors were represented in the sample
- A small sample of synthetic image data were included
- PHI was burned into the pixels in some images

MIDI	modality	patients	studies	series	instances
1.0	CR	32	41	42	45
1.0	CT	112	115	119	21,005
1.0	DX	53	56	56	71
1.0	MG	31	31	31	46
1.0	MR	103	105	138	7,452
1.0	PT	101	133	140	27,271

MIDI	modality	patients	studies	series	instances
1.1	CR	65	73	75	78
1.1	CT	120	147	150	14,517
1.1	DX	64	72	75	107
1.1	MG	74	74	75	90
1.1	MR	157	167	175	10,828
1.1	PT	88	117	150	42,187
1.1	SR	62	67	75	75
1.1	US	73	75	75	283

MIDI	modality	patients	studies	series	instances
1.2	MR	242	242	484	47,130
1.2	RTDOSE	242	242	484	484
1.2	RTPLAN	242	242	484	484
1.2	RTSTRUCT	242	242	484	484
1.2	WSI- Pathology	204			250

Examples of Unusual DICOM Data Elements Found to Contain PHI

- The table displays examples of DICOM data elements that should not contain PHI but have been found to do so during TCIA curation
- The frequency of occurrence identified in the analysis of TCIA audit logs is also indicated.

DICOM Tag	DICOM Description	Frequency
<(0008,0041)>	Data Set Subtype	1
<(0018,1250)>	Receive Coil Name	2
<(0018,7006)>	Detector Description	3
<(0010,0021)>	Issuer of Patient ID	4
<(0032,1030)>	Reason for Study	5
<(0008,1080)>	Admitting Diagnoses Description	6
<(0032,1000)>	Scheduled Study Start Date	11
<(0018,0010)>	Contrast/Bolus Agent	15
<(0018,1401)>	Acquisition Device Processing Code	29
<(0018,1000)>	Device Serial Number	31
<(0008,1010)>	Station Name	33
<(0032,1060)>	Requested Procedure Description	37
<(0008,2111)>	Derivation Description	44
<(3006,0006)>	Structure Set Description	50
<(3006,0008)>	Structure Set Date	57
<(0032,4000)>	Study Comments	70
<(0010,21b0)>	Additional Patient History	76
<(0032,1070)>	Requested Contrast Agent	101
<(0008,1030)>	Study Description	297
<(0010,4000)>	Patient Comments	1192

Example Private DICOM Attributes containing PHI.

- Examples of Private DICOM Data Elements, and their frequency of occurrence identified in the analysis of TCIA audit logs.
- Based on TCIA knowledgebase of Private Attributes

DICOM Tag	DICOM Description	Frequency
<(0027,"GEMS_IMAG_01",33)>	ImagingOptions	1
<(3f01,"INTELERAD MEDICAL SYSTEMS",03)>	SourceAE	1
<(7005,"TOSHIBA_MEC_CT3",1c)>	Contrast/Bolus Agent for Series Record	1
<(0009,"GEMS_PETD_01",37)>	Batch Description	2
<(0045,"GEMS_SENO_02",26)>	MAOBuffer	2
<(0009,"FDMS 1.0",92)>	KanjiDepartmentName	3
<(0009,"GEMS_IDEN_01",30)>	ServiceId	4
<(0043,"GEMS_PARM_01",80)>	Coil ID Data	8
<(0021,"SIEMENS MR SDS 01",19)>	MR Phoenix Protocol	15
<(0023,"GEMS_STDY_01",70)>	StartTimeSecsInFirstAxial	156

Evaluation Answer Key

Scope	Tag	Tag Name	Action	Action Text
<Study>	<(0008,0050)>	<Accession Number>	<text_removed>	<["20130830E626254"]>
<Study>	<(0008,1030)>	<Study Description>	<text_removed>	<["Stephanie Meyer"]>
<Study>	<(0008,1030)>	<Study Description>	<text_retained>	<["XR CHEST AP PORTABLE"]>
<Study>	<(0008,0080)>	<Institution Name>	<text_removed>	<["Dunn-Lindsey Memorial"]>
<Study>	<(0008,0090)>	<Referring Physician's Name>	<text_removed>	<["MORTON^JANET"]>
<Patient>	<(0010,0020)>	<Patient ID>	<text_removed>	<["8548156246"]>
<Patient>	<(0010,0010)>	<Patient's Name>	<text_removed>	<["MEYER^STEPHANIE"]>
<Patient>	<(0010,0030)>	<Patient's Birth Date>	<text_removed>	<["19530716"]>
<Patient>	<(0010,2154)>	<Patient's Telephone Numbers>	<text_removed>	<["+1-557-989-3970"]>
<Series>	<(0010,0010)>	<Patient's Name>	<tag_retained>	
<Series>	<(0010,0020)>	<Patient ID>	<tag_retained>	
<Series>	<(0010,0030)>	<Patient's Birth Date>	<tag_retained>	
<Series>	<(0010,0040)>	<Patient's Sex>	<tag_retained>	
<Series>	<(0020,000d)>	<Study Instance UID>	<tag_retained>	
<Series>	<(0020,000d)>	<Study Instance UID>	<text_notnull>	
<Series>	<(0008,1150)>	<Referenced SOP Class UID>	<text_retained>	<["1.2.840.10008.3.1.2.3.3"]>
<Series>	<(0008,0016)>	<SOP Class UID>	<text_retained>	<["1.2.840.10008.5.1.4.1.1.1.1"]>
<Series>	<(0008,1155)>	<Referenced SOP Instance UID>	<uid_changed>	<["2.25.160539642186938793107880005813476638198"]>
<Series>	<(0020,000e)>	<Series Instance UID>	<uid_changed>	<["2.25.267046995551041197882743054101055651318"]>
<Series>	<(0008,0018)>	<SOP Instance UID>	<uid_changed>	<["2.25.56598319458405629701989395893226946038"]>
<Series>	<(0020,000d)>	<Study Instance UID>	<uid_changed>	<["2.25.66491365017322661013109147239553023478"]>
<Series>	<(0008,0020)>	<Study Date>	<date_shifted>	<["20130829","(552)"]>
<Series>	<(0008,0021)>	<Series Date>	<date_shifted>	<["20130829","(552)"]>
<Instance>			<pixels_hidden>	<["MEYER STEPHANIE [F] 02.25.2012 \\n DOB: 07.16.1953", [0, 0, 450, 150,48]]>
<Instance>	<(0008,0030)>	<Study Time>	<text_retained>	<["191128"]>
<Instance>	<(0008,0031)>	<Series Time>	<text_retained>	<["191130.000000"]>
<Instance>	<(0008,0032)>	<Acquisition Time>	<text_retained>	<["191131"]>
<Instance>	<(0008,0033)>	<Content Time>	<text_retained>	<["191131"]>

Action	Description
tag_retained	The tag itself is retained and present in the DICOM dataset
text_notnull	The value of the tag is not null or zero length value
text_retained	The text specified was retained in the tag value
text_removed	The test specified was removed from the tag value
date_shifted	The date was shifted using the specified shift value
uid_changed	The UID was updated according to curation crosswalk
pixels_hidden	The pixels within coordinates specified are hidden