Medical Image De-Identification (MIDI) Task Group (TG)

MIDI Workshop – 2023/05/22

David A. Clunie, PixelMed

Support

- This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. 75N91019D00024, Task Order 75N91019F00129. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government

Acknowledgements

- Leadership
- Task Group members
- Steering Committee members
- External reviewers
- Staff

- See report for complete list

Task Group Members

Adam Flanders	Thomas Jefferson University
Adam Taylor	Sage Bionetworks
Brad Erickson	Mayo Clinic
Brian Bialecki	American College of Radiology
David Brundage	Cornell University
David Clunie	PixelMed Publishing
David Gutman	Emory University
Fred Prior	University of Arkansas for Medical Sciences
J Anthony Seibert	University of California, Davis
John Perry	Independent consultant
Judy Wawira Gichoya	Emory University
Justin Kirby	Frederick National Laboratory for Cancer Research
Katherine Andriole	Brigham and Women's Hospital
Luke Geneslaw	Memorial Sloan Kettering Cancer Center
Steve Moore	Washington University in St. Louis
TJ Fitzgerald	UMass Memorial Medical Center
Wyatt Tellis	University of California, San Francisco
Ying Xiao	University of Pennsylvania Health System

Background

- Keyvan Farahani initiated overall MIDI (initiative) spring 2020
  - datasets
  - pipeline
  - task group on best practices
  - …
- TG initiated Jun 2021
- Convened virtually, monthly, until Nov 2022, 15 times
- Solicited outside input prn
- Report V1 issued as pre-print on arXiv Mar 2023
- Plan to publish as journal supplement and/or executive summary

MIDI Task Group Mission, Goals, Charge

- To document strategies and best practices in medical image de-ID for secondary sharing of imaging data with an emphasis on DICOM
- To reach consensus on best practices
- To disseminate findings
- To provide input toward CBIIT/NCI and other ICs activities
- To make recommendations on criteria and resources for performance evaluation of tools
- To provide guidelines for image de-ID using automated vs. manual, cloud-based vs. local approaches, portability, scalability

Scope

- Medical images of human subjects and biospecimens
- Re-identification risk sufficiently reduced for unrestricted public sharing for any purpose
- All medical images, regardless of the mode of acquisition
  - including anatomical pathology Whole Slide Imaging (WSI)
- Also related non-image objects, such as:
  - RT Structure Sets, Plans and Dose Volume Histograms
  - Structured Reports and Presentation States
- Particularly, but not only, DICOM

Deliverable - Report

- Best Practices
  - what you should be doing now
- Recommendations
  - further research, investigation, development, documentation
- Comprehensive
  - approximately 80 pages of text + 46 pages of references
  - 18 best practices and 8 recommendations

Methodology

- Extensive literature review
  - informal: not a systematic review
  - searches
  - citation tracking
  - suggestions from members and reviewers
- Discuss major, difficult, unexplored, controversial topics (monthly)
  - traditional methods – standard rule based approaches to structured metadata
  - burned in text recognition and redaction
  - Potentially Reconstructable Facial Information (PRFI)
  - threat models and quantification of re-identification risk
  - Statistical Disclosure Control (SDC) and lessons from microdata community
  - use of automated approaches including AI

Overview of Content

- Best Practices and Recommendations
- Scope
- Terminology
- File formats – DICOM, non-DICOM, standard and proprietary, private extensions
- What needs to be de-identified – within files, in accompanying or linked data sets (e.g., clinical)
- Rule-based de-identification (emphasis on DICOM PS3.15 profile)
- Statistical Disclosure Control (SDC): re-identification threat model, risk, indirect identifiers +/- modification
- Structured, unstructured, burned-in, dates
- Image features – derivation of face, age, sex, race from photos, radiography
- Metadata lurking in obscure places – inside JPEG bitstream
- Modality-specific issues – including external photos, WSI
- AI used for de-identification (not just as customer for de-identified data)
- Reports, documents, annotations
- Evaluation, scoring, motivated intruder attack
- Operational and deployment considerations, including scalability, quality control, tools

http://says.com/my/lifestyle/what-you-should-do-with-your-salary

Best Practice #1 - Everything & quantify risk

- *"Thorough de-identification by removal or replacement of all known direct and indirect identifiers and sensitive information, in all collection descriptions and supporting data, structured and unstructured text data elements, pixel data, and geometric and bitmapped overlays, is required for public sharing. Direct identifiers should always be removed. A realistic collection-specific expert statistical analysis should be performed to quantify residual re-identification risk with respect to a pre-determined risk threshold, to justify retention of selected indirect identifiers or sensitive information, potentially with modified risk-reducing values, to preserve re-use utility. Any such risk analysis needs to consider any other publicly available information about the subject, and is only valid at the point in time at which it was done; consideration should be given to the potential for an increase in risk over time."*

Best Practice #1 - Everything & quantify risk

- *"Thorough de-identification by removal or replacement of all known direct and indirect identifiers and sensitive information, in all collection descriptions and supporting data, structured and unstructured text data elements, pixel data, and geometric and bitmapped overlays, is required for public sharing. Direct identifiers should always be removed. A realistic collection-specific expert statistical analysis should be performed to quantify residual re-identification risk with respect to a pre-determined risk threshold, to justify retention of selected indirect identifiers or sensitive information, potentially with modified risk-reduced values, to preserve re-use utility. Any such risk analysis needs to consider any other publicly available information about the subject, and is only valid at the point in time at which it was done; consideration should be given to the potential for an increase in risk over time."*

ARX

Data Anonymization Tool

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data.

The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes.

ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface. You can find further information here, or directly proceed to our downloads section.

https://arx.deidentifier.org/

ARX Anonymization Tool - MIDI_TG_Experiment_CPTAC

Attribute: age Transformations: 400 Selected: [0, 0, 0] Applied: [0, 0, 0]

Configure transformation | Explore results | Analyze utility | Analyze risk

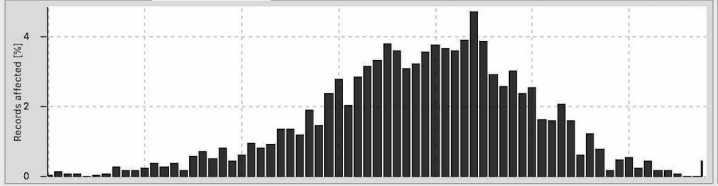
Input data | Classification performance | Quality models

	slide_id	specimen_id	tumor_code	case_id	gender	
1	f3cf4b6c-c7b...	e9ac0447-65...	BR	01BR030	Female	84
2	e9ac0447-65...	e9ac0447-65...	BR	01BR030	Female	84
3	6e598525-34...	e9ac0447-65...	BR	01BR030	Female	84
4	1f5a9aad-61a...	e9ac0447-65...	BR	01BR030	Female	84
5	C3N-01802-26	C3N-01802-06	UCEC	C3N-01802	Female	85
6	C3N-01802-21	C3N-01802-01	UCEC	C3N-01802	Female	85
7	C3N-02259-21	C3N-02259-01	GBM	C3N-02259	Female	84
8	C3N-01874-25	C3N-01874-05	UCEC	C3N-01874	Female	84
9	C3N-01874-23	C3N-01874-03	UCEC	C3N-01874	Female	84
10	C3N-01874-21	C3N-01874-01	UCEC	C3N-01874	Female	84
11	C3N-03415-21	C3N-03415-01	UCEC	C3N-03415	Female	85
12	C3L-04853-24	C3L-04853-04	PDA	C3L-04853	Female	85
13	C3L-04853-23	C3L-04853-03	PDA	C3L-04853	Female	85
14	C3L-04853-22	C3L-04853-02	PDA	C3L-04853	Female	85
15	C3L-04853-21	C3L-04853-01	PDA	C3L-04853	Female	85
16	C3N-02723-26	C3N-02723-06	CCRCC	C3N-02723	Female	86
17	C3N-02723-23	C3N-02723-03	CCRCC	C3N-02723	Female	86
18	C3N-02723-22	C3N-02723-02	CCRCC	C3N-02723	Female	86
19	C3N-02723-21	C3N-02723-01	CCRCC	C3N-02723	Female	86
20	b10c7b72-1d4...	[50]-f63649_...	BR	11BR054	Female	84
21	C3L-02604-21	C3L-02604-01	PDA	C3L-02604	Female	84
22	C3L-01732-22	C3L-01732-02	UCEC	C3L-01732	Female	84
23	C3L-01732-21	C3L-01732-01	UCEC	C3L-01732	Female	84
24	C3L-01063-26	C3L-01063-06	CM	C3L-01063	Female	84

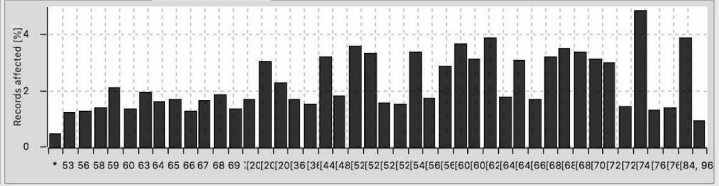
Output data | Classification performance | Quality models

	slide_id	specimen_id	tumor_code	case_id	gender	
1	f3cf4b6c-c7b...	e9ac0447-65...	BR	01BR030	*	[20, 96
2	e9ac0447-65...	e9ac0447-65...	BR	01BR030	*	[20, 96
3	6e598525-34...	e9ac0447-65...	BR	01BR030	*	[20, 96
4	1f5a9aad-61a...	e9ac0447-65...	BR	01BR030	*	[20, 96
5	C3N-01802-26	C3N-01802-06	UCEC	C3N-01802	*	[20, 96
6	C3N-01802-21	C3N-01802-01	UCEC	C3N-01802	*	[20, 96
7	C3N-02259-21	C3N-02259-01	GBM	C3N-02259	*	[20, 96
8	C3N-01874-25	C3N-01874-05	UCEC	C3N-01874	*	[20, 96
9	C3N-01874-23	C3N-01874-03	UCEC	C3N-01874	*	[20, 96
10	C3N-01874-21	C3N-01874-01	UCEC	C3N-01874	*	[20, 96
11	C3N-03415-21	C3N-03415-01	UCEC	C3N-03415	*	[20, 96
12	C3L-04853-24	C3L-04853-04	PDA	C3L-04853	*	[20, 96
13	C3L-04853-23	C3L-04853-03	PDA	C3L-04853	*	[20, 96
14	C3L-04853-22	C3L-04853-02	PDA	C3L-04853	*	[20, 96
15	C3L-04853-21	C3L-04853-01	PDA	C3L-04853	*	[20, 96
16	C3N-02723-26	C3N-02723-06	CCRCC	C3N-02723	*	[20, 96
17	C3N-02723-23	C3N-02723-03	CCRCC	C3N-02723	*	[20, 96
18	C3N-02723-22	C3N-02723-02	CCRCC	C3N-02723	*	[20, 96
19	C3N-02723-21	C3N-02723-01	CCRCC	C3N-02723	*	[20, 96
20	b10c7b72-1d4...	[50]-f63649_...	BR	11BR054	*	[20, 96
21	C3L-02604-21	C3L-02604-01	PDA	C3L-02604	*	[20, 96
22	C3L-01732-22	C3L-01732-02	UCEC	C3L-01732	*	[20, 96
23	C3L-01732-21	C3L-01732-01	UCEC	C3L-01732	*	[20, 96
24	C3L-01063-26	C3L-01063-06	CM	C3L-01063	*	[20, 96

Summary statistics | Distribution | Contingency | Class sizes | Properties



Summary statistics | Distribution | Contingency | Class sizes | Properties

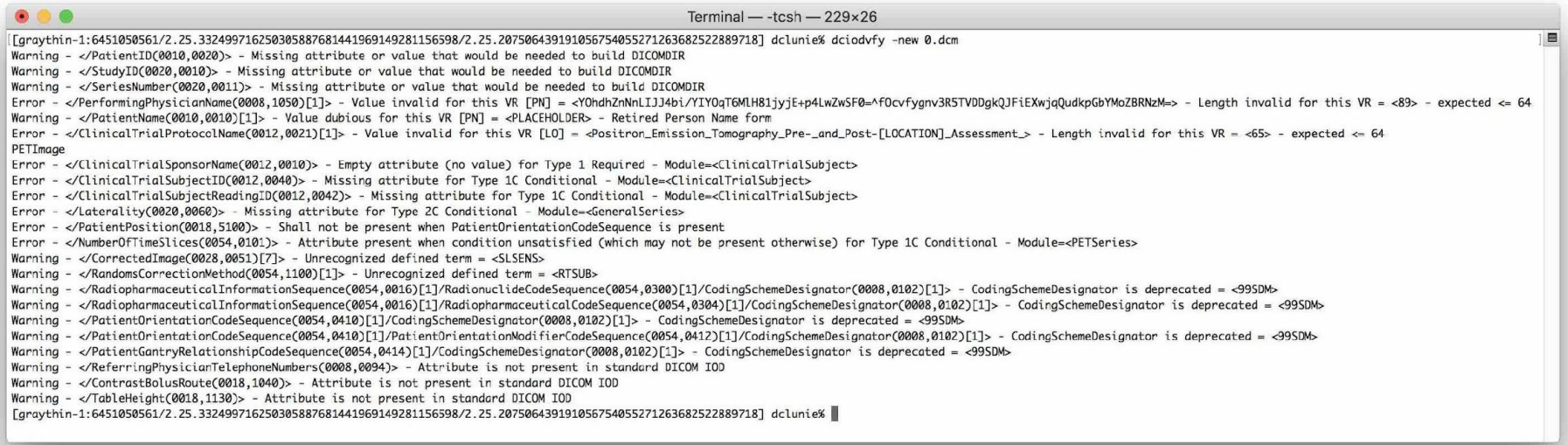


Best Practice #3 - Remain compliant

- *"The de-identification process should not compromise the conformance of the resulting data with the standards that define the content, or reduce the level of functionality; specifically, de-identification of DICOM files should retain DICOM conformance with the original information object definition (IOD), even if that requires synthesis of dummy values for replacement, and consistent replacement values across multiple files (e.g., to retain referential integrity of replaced unique identifiers within a defined scope). This requires retention or replacement of not only required attributes, but also optional attributes critical to retain functionality."*

Best Practice #3 - Remain compliant

- *"The de-identification process should not compromise the conformance of the resulting data with the standards that define the content, or reduce the level of functionality; specifically, de-identification of DICOM files should retain DICOM conformance with the original information object definition (IOD), even if that requires synthesis of dummy values for replacement, and consistent replacement values across multiple files (e.g., to retain referential integrity of replaced unique identifiers within a defined scope). This requires retention or replacement of not only required attributes, but also optional attributes critical to retain functionality."*



http://tinyurl.com/dciodvfy

Best Practice #4 - Preserve utility

- *"The de-identification process should preserve as much information about the image acquisition as possible (including machine identity, characteristics, and settings) to maximize the re-use potential, except to the extent that machine information can be realistically quantified as increasing the residual re-identification risk above a pre-determined acceptable risk threshold."*

Best Practice #4 - Preserve utility

- *"The de-identification process should preserve as much information about the image acquisition as possible (including machine identity, characteristics, and settings) to maximize the re-use potential, except to the extent that machine information can be realistically quantified as increasing the residual re-identification risk above a pre-determined acceptable risk threshold."*

Best Practice #6 - Use the standard profile

- *"For DICOM images, the current release … of the DICOM PS3.15 E.1 Application Level Confidentiality Profile should be used as a reference for those structured and unstructured data elements that need to be de-identified, augmented by any additional knowledge of other unsafe attributes, including private data elements, that need to be considered …The PS3.15 approach of removing or replacing everything that is known to be unsafe, and retaining only what is known to be safe … is applicable to any DICOM object, whether an image or not ... various options beyond the baseline for retention, cleaning, or removal of information for various scenarios, and these choices should be carefully evaluated to balance preservation of utility against residual re-identification risk ..."*

Best Practice #6 - Use the standard profile

- *"For DICOM images, the current release … of the DICOM PS3.15 E.1 Application Level Confidentiality Profile should be used as a reference for those structured and unstructured data elements that need to be de-identified, augmented by any additional knowledge of other unsafe attributes, including private data elements, that need to be considered …The PS3.15 approach of removing or replacing everything that is known to be unsafe, and retaining only what is known to be safe … is applicable to any DICOM object, whether an image or not ... various options beyond the baseline for retention, cleaning, or removal of information for various scenarios, and these choices should be carefully evaluated to balance preservation of utility against residual re-identification risk ..."*

DICOM PS3.15 Annex E – Problem Statement

- Nobody does it ~~better~~ ***the same way***
- Regulatory requirements vary by jurisdiction, change over time
- Ethical, moral, legal requirements
- Lack of disciplined risk analysis
- Inexperience
- Complexity of problem, encoding, etc.
- Lack of consistent tools
- Left to the user

DICOM PS3.15 Annex E – Solution

- Extend DICOM standard to address the issue properly
- Sup 142 Jan 2011 was added to DICOM PS3.15:
  - *"to provide instruction for implementers,*
  - *to assure compliance, and*
  - *to provide guidance for sites and trial administrators that has been subject to expert review"*
- Tabulate ALL potential ID-containing attributes and deal with them
- Specific options to remove or retain more depending on utility reqd.
- **Maintained** as new attributes added to DICOM for new use cases
- Machine-readable – use XML tables to automagically update tools

Table E.1-1a. De-identification Action Codes

D	replace with a non-zero length value that may be a dummy value and consistent with the VR
Z	replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
X	remove
K	keep (unchanged for non-Sequence Attributes, cleaned for Sequences)
C	clean, that is replace with values of similar meaning known not to contain identifying information and consistent with the VR
U	replace with a non-zero length UID that is internally consistent within a set of Instances
Z/D	Z unless D is required to maintain IOD conformance (Type 2 versus Type 1)
X/Z	X unless Z is required to maintain IOD conformance (Type 3 versus Type 2)
X/D	X unless D is required to maintain IOD conformance (Type 3 versus Type 1)
X/Z/D	X unless Z or D is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1)
X/Z/U*	X unless Z or replacement of contained instance UIDs (U) is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1 sequences containing UID references)

Table E.1-1. Application Level Confidentiality Profile Attributes

Attribute Name	Tag	Retd. (from PS3.6)	In Std. Comp. IOD (from PS3.3)	Basic Prof.	Rtn. Safe Priv. Opt.	Rtn. UIDs Opt.	Rtn. Dev. Id. Opt.	Rtn. Inst. Id. Opt.	Rtn. Pat. Chars. Opt.	Rtn. Long. Full Dates Opt.	Rtn. Long. Modif. Dates Opt.	Clean Desc. Opt.	Clean Struct. Cont. Opt.	Clean Graph. Opt.
Accession Number	(0008,0050)	N	Y	Z										
Acquisition Comments	(0018,4000)	Y	N	X								C		
Acquisition Context Sequence	(0040,0555)	N	Y	X/Z									C	
Acquisition Date	(0008,0022)	N	Y	X/Z						K	C			
Acquisition DateTime	(0008,002A)	N	Y	X/Z/D						K	C			
Acquisition Device Processing Description	(0018,1400)	N	Y	X/D								C		
Acquisition Field Of View Label	(0018,11BB)	N	Y	D								C		
Acquisition Protocol Description	(0018,9424)	N	Y	X								C		
Acquisition Time	(0008,0032)	N	Y	X/Z						K	C			
Acquisition UID	(0008,0017)	N	Y	U		K								
Actual Human Performers Sequence	(0040,4035)	N	N	X										
Additional Patient History	(0010,21B0)	N	Y	X								C		

Baseline – Extremely conservative approach

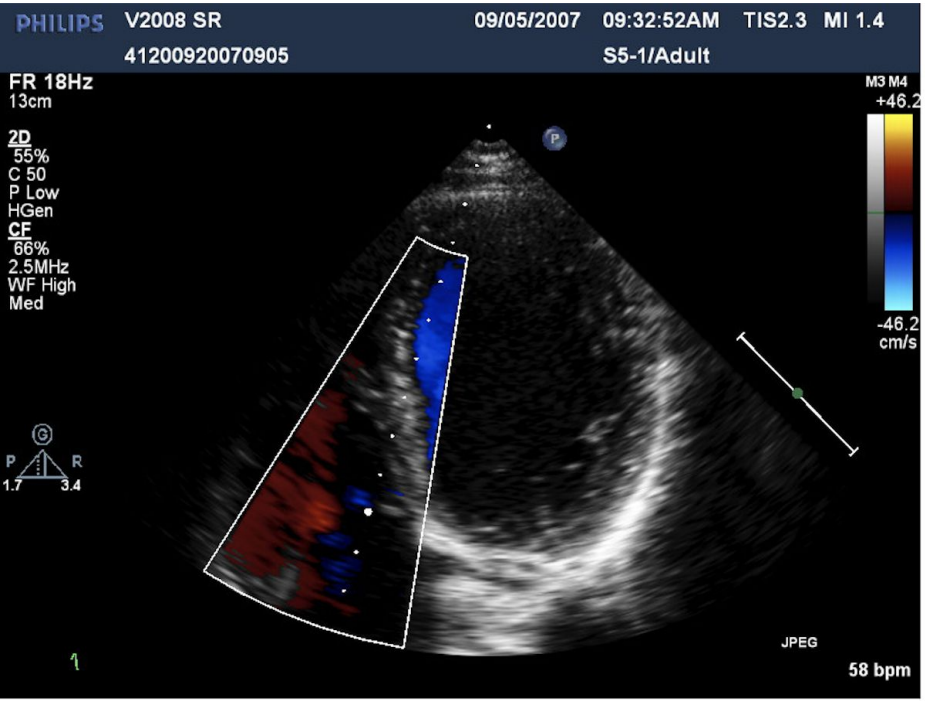
- Removes all information related to:
  - identity and demographic characteristics of the patient
  - identity of any responsible parties or family members
  - identity of any personnel involved in the procedure
  - identity of the organizations involved in ordering/performing procedure
  - additional information that could be used to match instances if given access to the originals, such as UIDs, dates and times
  - private data elements

Table E.1-1. Application Level Confidentiality Profile Attributes

Attribute Name	Tag	Retd. (from PS3.6)	In Std. Comp. IOD (from PS3.3)	Basic Prof.	Rtn. Safe Priv. Opt.	Rtn. UIDs Opt.	Rtn. Dev. Id. Opt.	Rtn. Inst. Id. Opt.	Rtn. Pat. Chars. Opt.	Rtn. Long. Full Dates Opt.	Rtn. Long. Modif. Dates Opt.	Clean Desc. Opt.	Clean Struct. Cont. Opt.	Clean Graph. Opt.
Accession Number	(0008,0050)	N	Y	Z										
Acquisition Comments	(0018,4000)	Y	N	X								C		
Acquisition Context Sequence	(0040,0555)	N	Y	X/Z									C	
Acquisition Date	(0008,0022)	N	Y	X/Z						K	C			
Acquisition DateTime	(0008,002A)	N	Y	X/Z/D						K	C			
Acquisition Device Processing Description	(0018,1400)	N	Y	X/D								C		
Acquisition Field Of View Label	(0018,11BB)	N	Y	D								C		
Acquisition Protocol Description	(0018,9424)	N	Y	X								C		
Acquisition Time	(0008,0032)	N	Y	X/Z						K	C			
Acquisition UID	(0008,0017)	N	Y	U		K								
Actual Human Performers Sequence	(0040,4035)	N	N	X										
Additional Patient History	(0010,21B0)	N	Y	X								C		

Options – Remove, Process

- Removal of additional information:
  - Clean Pixel Data
  - Clean Recognizable Visual Features
  - Clean Graphics
- Processing if otherwise removed but needed for specific uses:
  - Clean Structured Content
  - Clean Descriptors (i.e., plain text, short or long)


PHILIPS
V2008 SR
09/05/2007 09:32:52AM TIS2.3 MI 1.4
41200920070905
S5-1/Adult
FR 18Hz
13cm
2D
55%
C 50
P Low
HGen
CF
66%
2.5MHz
WF High
Med
M3 M4
+46.2
-46.2
cm/s
P
R
1.7
3.4
JPEG
58 bpm

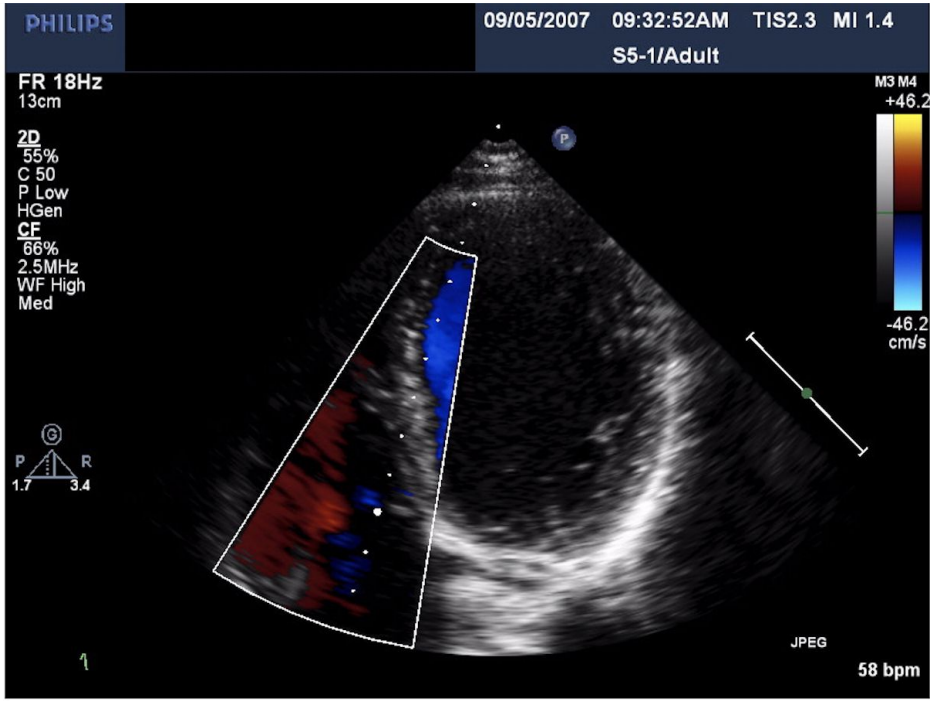

PHILIPS
09/05/2007 09:32:52AM TIS2.3 MI 1.4
S5-1/Adult
FR 18Hz
13cm
2D
55%
C 50
P Low
HGen
CF
66%
2.5MHz
WF High
Med
M3 M4
+46.2
-46.2
cm/s
P
R
1.7
3.4
JPEG
58 bpm

Table E.3.4-1. Application Level Confidentiality Profile Clean Structured Content Option Content Item Concept Name Codes

Code Meaning	Code Value	Coding Scheme Designator	Value Type	Retd. (from PS3.16)	In Std. Tpl. (from PS3.16)	Basic Prof.	Rtn. UIDs Opt.	Rtn. Dev. Id. Opt.	Rtn. Inst. Id. Opt.	Rtn. Pat. Chars. Opt.	Rtn. Long. Full Dates Opt.	Rtn. Long. Modif. Dates Opt.	Clean Desc. Opt.
Accession Number	121022	DCM	TEXT	N	Y	X							
Acquired Image	113795	DCM	IMAGE	N	Y	D	K						
Acquisition Date	126201	DCM	DATE	N	Y	X					K	C	
Acquisition Protocol	125203	DCM	TEXT	N	Y	X							C
Acquisition Time	126202	DCM	TIME	N	Y	X					K	C	
Activity Session	C67447	NCIt	TEXT	N	Y	X							C
Administration of radiopharmaceutical	440252007	SCT	TEXT	N	Y	D							C
Admission DateTime	15	NCDR [2.0b]	DATETIME	N	Y	X					K	C	
Anatomic Identifier	112050	DCM	TEXT	N	Y	X							C
Anesthesia Finish Time	398164008	SCT	DATETIME	N	Y	X					K	C	
Anesthesia Start Time	398325003	SCT	DATETIME	N	Y	X					K	C	
Best illustration of finding	121080	DCM	IMAGE	N	Y	X	K						
Best illustration of finding	121080	DCM	WAVEFORM	N	Y	X	K						
Calibration DateTime	113723	DCM	DATETIME	N	Y	D					K	C	
Calibration Protocol	113720	DCM	TEXT	N	Y	X							C

Clean Descriptors

- E.g., DICOM Study Description attribute
  - high utility preservation value
  - unstructured content
  - under scanner operator control & frequently abused for comments
- Before:

  CT Chest/Abdo/Pelvis per Dr. Smith protocol – call 212-555-1212

- After:

  CT Chest/Abdo/Pelvis per protocol – call

Clean Descriptors

- E.g., DICOM Study Description attribute
  - high utility preservation value
  - unstructured content
  - under scanner operator control & frequently abused for comments
- Before:

  MR Hand per Dr. Hand protocol

- After:

  MR Hand per protocol

Confidentiality Options - Retain

- Removal of additional information:
  - Retain Longitudinal Temporal Information
    - Full Dates
    - Modified Dates
  - Retain Patient Characteristics
  - Retain Device Identity
  - Retain Institution Identity
  - Retain UIDs
  - Retain Safe Private

PS3.15 – No Excuses – Accept No Less

- All de-identification "products" should implement PS3.15+/-Options
- May then do less or more
  - more – e.g., Age > 89Y top coded per HIPAA Privacy Rule 18 elements
  - less – must be justified with a risk analysis and ***documented***
- Extremely configurable products need to be supplied with "template" that uses ***current*** release of PS3.15 + has maintenance method
- Do not expect the user to be responsible for this
- With great power comes great ~~responsibility~~ ***opportunity for disaster***
- Use an industry-standard like PS3.15 to mitigate risk

Best Practice #8 - Non-DICOM

- *"For non-DICOM images, in the absence of an alternative specific reliable reference for data element retention or removal, the general principles explicit or implicit in DICOM PS3.15 E.1 should be applied, e.g., for images stored in DICOM-derived formats like Brain Imaging Data Structure (BIDS) with an alternative metadata representation. For clinical data elements, the general principles in the PhUSE De-Identification Standard for CDISC SDTM should be applied."*

Best Practice #8 - Non-DICOM

- *"For non-DICOM images, in the absence of an alternative specific reliable reference for data element retention or removal, the general principles explicit or implicit in DICOM PS3.15 E.1 should be applied, e.g., for images stored in DICOM-derived formats like Brain Imaging Data Structure (BIDS) with an alternative metadata representation. For clinical data elements, the general principles in the PhUSE De-Identification Standard for CDISC SDTM should be applied."*

Best Practice #9 - All elements anywhere

- *"Regardless of the image encoding or file format, all data elements linked to images in the collection, including those in accompanying spreadsheets or publications, which are linked by a common key (e.g., the pseudonymous subject identifier) need to be de-identified and subject to a risk analysis. That risk analysis should account for linked information in other public data sets for the same subjects, which are made available by other organizations and that are known to the de-identifier ... A search for the existence of such linked data should be undertaken."*

Best Practice #9 - All elements anywhere

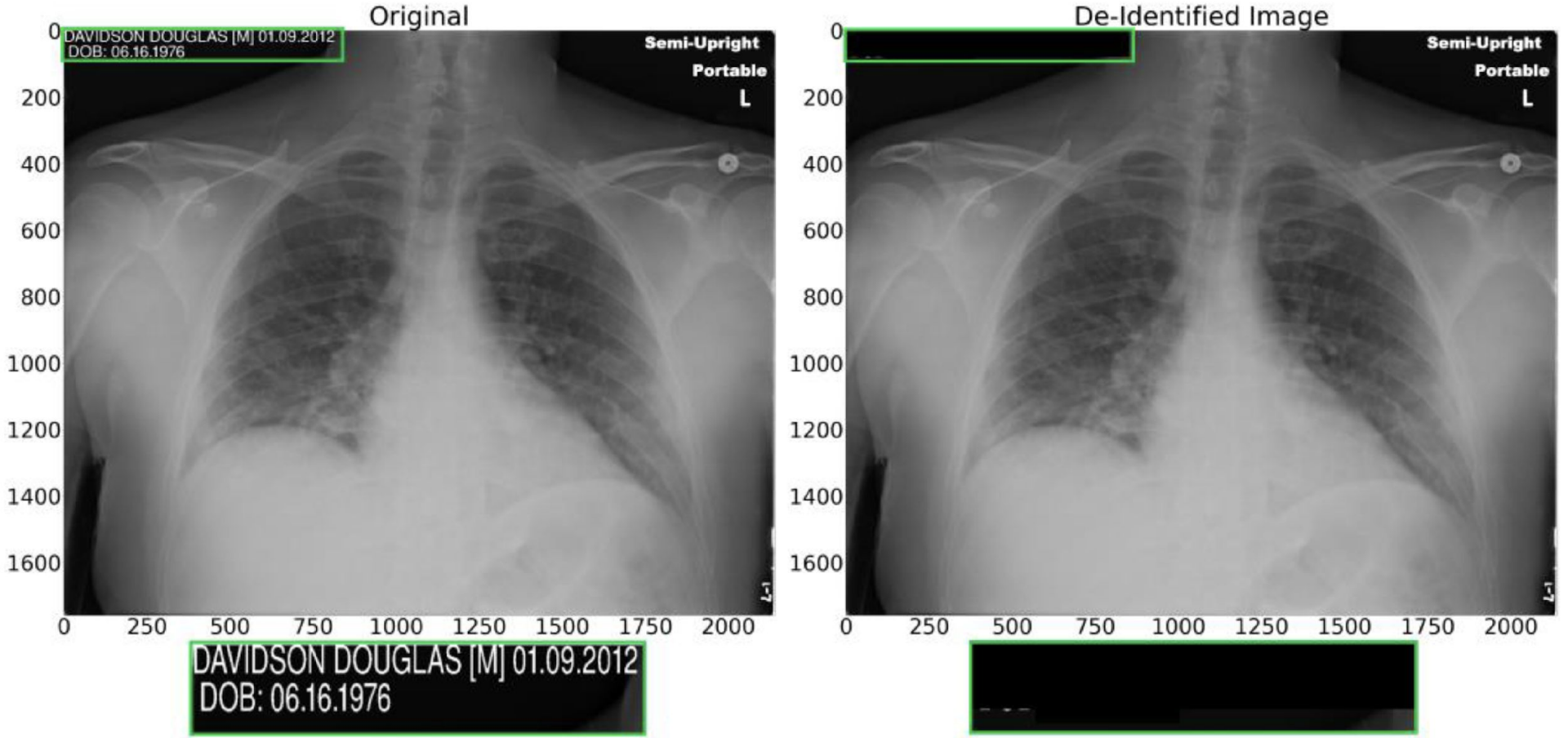
- *"Regardless of the image encoding or file format, all data elements linked to images in the collection, including those in accompanying spreadsheets or publications, which are linked by a common key (e.g., the pseudonymous subject identifier) need to be de-identified and subject to a risk analysis. That risk analysis should account for linked information in other public data sets for the same subjects, which are made available by other organizations and that are known to the de-identifier ... A search for the existence of such linked data should be undertaken."*

Best Practice #10 - Burned-in text

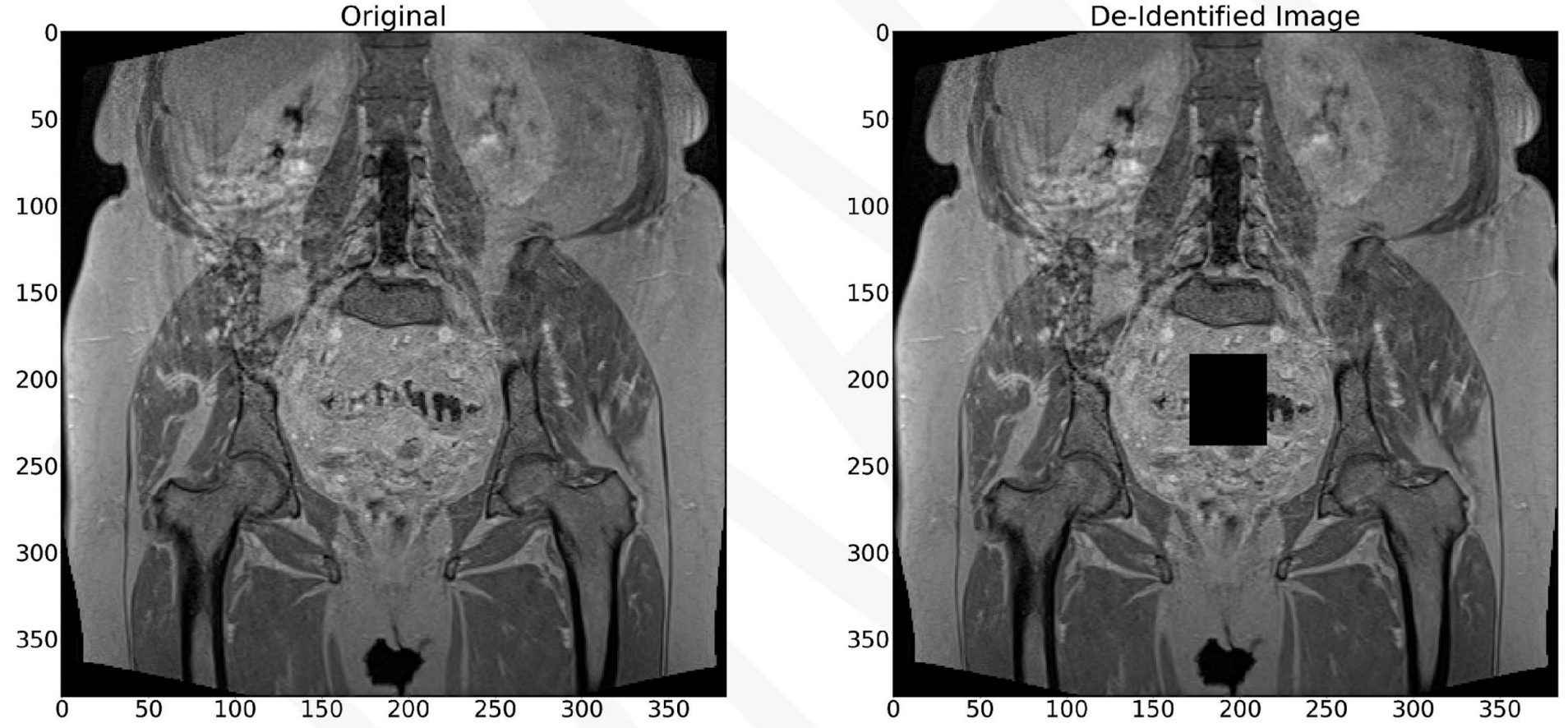
- *"The risk posed by the presence of burned-in text, foreign objects with textual information (e.g., jewelry) and other sources of potential identity leakage in pixel data should be assessed, and if the risk exceeds a pre-determined threshold, scanned for the offending information, and the entire image discarded or the offending information redacted, manually or automatically (subject to subsequent human review); the effort to scan and redact versus discard should be weighed against re-use utility. This risk assessment should be performed for all image types ... It is not sufficient to limit checks for offending information to only a stratified sub-set of image types ..."*

Best Practice #10 - Burned-in text

- *"The risk posed by the presence of burned-in text, foreign objects with textual information (e.g., jewelry) and other sources of potential identity leakage in pixel data should be assessed, and if the risk exceeds a pre-determined threshold, scanned for the offending information, and the entire image discarded or the offending information redacted, manually or automatically (subject to subsequent human review); the effort to scan and redact versus discard should be weighed against re-use utility. This risk assessment should be performed for all image types ... It is not sufficient to limit checks for offending information to only a stratified sub-set of image types ..."*



Kopchick et al. Medical image de-identification using cloud services. SPIE MI 2022.



Kopchick et al. MIDI Pilot Final Report. 2021. Internal CBIIT Report from Deloitte

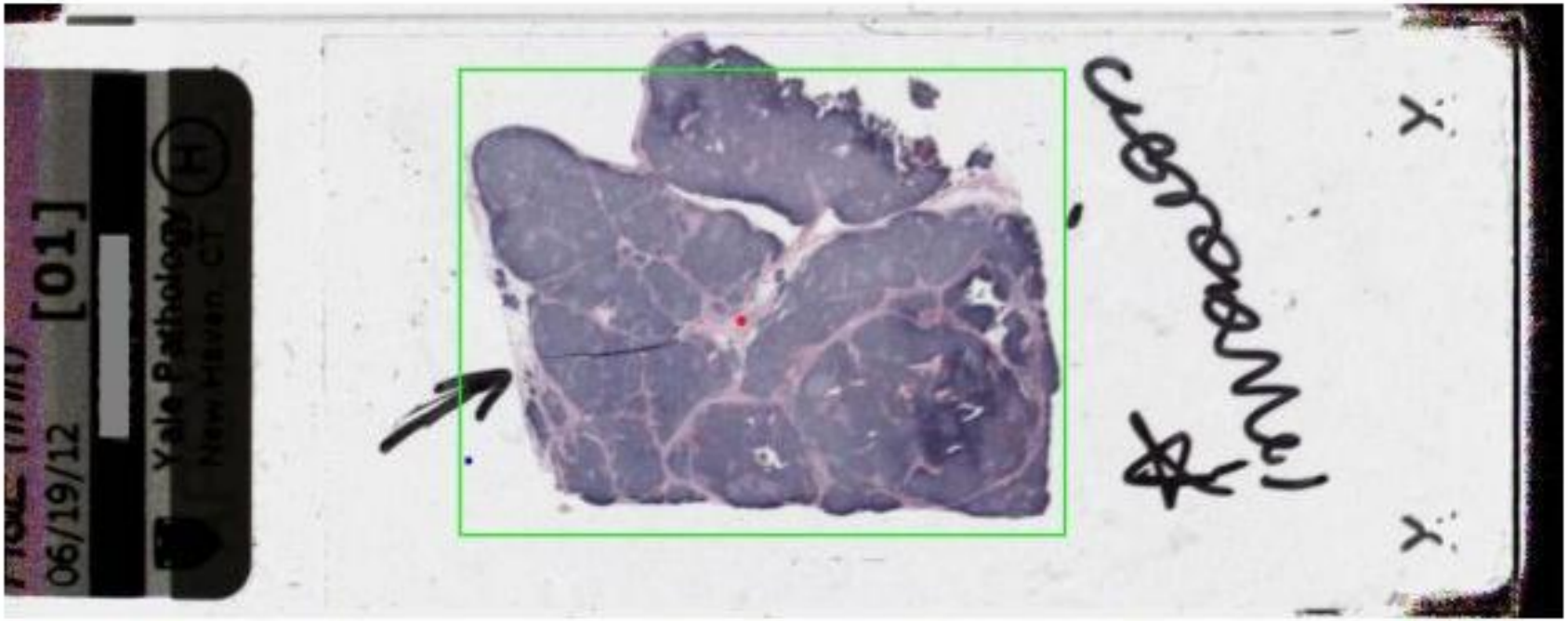


Figure 4: The macro image from TCGA file TCGA-4X-A9FB-01Z-00-DX1.211CC9AA-F721-4D16-8663-68A393223F80.svs. The left side shows part of the label which included the patient name. This has been manually redacted with a grey bar in this figure. This file was released to the public in 2016.

Best Practice #13 - Private elements

- *"Private data elements retained to preserve utility should be evaluated with respect to risk of identity leakage, either by reference to a reliable source of known safe private data elements, such as that provided in DICOM PS3.15 E.3.10, manufacturer's documentation, including DICOM Conformance Statements, or published documents from other reliable sources. Otherwise, private data elements should be selectively or entirely removed."*

Best Practice #13 - Private elements

- *"Private data elements retained to preserve utility should be evaluated with respect to risk of identity leakage, either by reference to a reliable source of known safe private data elements, such as that provided in DICOM PS3.15 E.3.10, manufacturer's documentation, including DICOM Conformance Statements, or published documents from other reliable sources. Otherwise, private data elements should be selectively or entirely removed."*

Table E.3.10-1. Safe Private Attributes

Data Element	Private Creator	VR	VM	Meaning
(7053,xx00)	Philips PET Private Group	DS	1	SUV Factor - Multiplying Stored Pixel Values by Rescale Slope then this factor results in SUVbw in g/l
(7053,xx09)	Philips PET Private Group	DS	1	Activity Concentration Factor - Multiplying Stored Pixel Values by Rescale Slope then this factor results in MBq/ml.
(00E1,xx21)	ELSCINT1	DS	1	DLP
(00E1,xx50)	ELSCINT1	DS	1	Acquisition Duration
(01E1,xx26)	ELSCINT1	CS	1	Phantom Type
(01F1,xx01)	ELSCINT1	CS	1	Acquisition Type
(01F1,xx07)	ELSCINT1	DS	1	Table Velocity
(01F1,xx26)	ELSCINT1	DS	1	Pitch
(01F1,xx27)	ELSCINT1	DS	1	Rotation Time
(0019,xx23)	GEMS_ACQU_01	DS	1	Table Speed [mm/rotation]
(0019,xx24)	GEMS_ACQU_01	DS	1	Mid Scan Time [sec]
(0019,xx27)	GEMS_ACQU_01	DS	1	Rotation Speed (Gantry Period)
(0019,xx9E)	GEMS_ACQU_01	LO	1	Internal Pulse Sequence Name
(0043,xx27)	GEMS_PARM_01	SH	1	Scan Pitch Ratio in the form "n.nnn:1"
(0045,xx01)	GEMS_HELIOS_01	SS	1	Number of Macro Rows in Detector
(0045,xx02)	GEMS_HELIOS_01	FL	1	Macro width at ISO Center
(0903,xx10)	GEIIS PACS	US	1	Reject Image Flag
(0903,xx11)	GEIIS PACS	US	1	Significant Flag
(0903,xx12)	GEIIS PACS	US	1	Confidential Flag
(2001,xx01)	Philips Imaging DD 001	FL	1	MR Image Chemical Shift
(2001,xx02)	Philips Imaging DD 001	IS	1	MR Image Chemical Shift Number
(2001,xx03)	Philips Imaging DD 001	FL	1	MR Image Diffusion B-Factor

Best Practice #14 - Obscure metadata

- *"Compressed bitstreams used as pixel data or within other data elements … should be considered with respect to the potential for identity leakage through embedded data elements, and either decompressed during de-identification (if losslessly compressed) and the embedded data elements discarded, or if the compressed bitstream is re-used, scanned for data elements at risk and those selectively removed or replaced. E.g., an EXIF APP1 or JUMBF APP11 marker segment in the lossy JPEG pixel data of a DICOM image may contain direct or indirect identifiers in data elements as well as information of re-use utility."*

Best Practice #14 - Obscure metadata

- *"Compressed bitstreams used as pixel data or within other data elements … should be considered with respect to the potential for identity leakage through embedded data elements, and either decompressed during de-identification (if losslessly compressed) and the embedded data elements discarded, or if the compressed bitstream is re-used, scanned for data elements at risk and those selectively removed or replaced. E.g., an EXIF APP1 or JUMBF APP11 marker segment in the lossy JPEG pixel data of a DICOM image may contain direct or indirect identifiers in data elements as well as information of re-use utility."*

Metadata takes **34.8 KB (0.5%)** of this image and **includes location data.** To protect your privacy, download this image without metadata by clicking the button below.

REMOVE METADATA

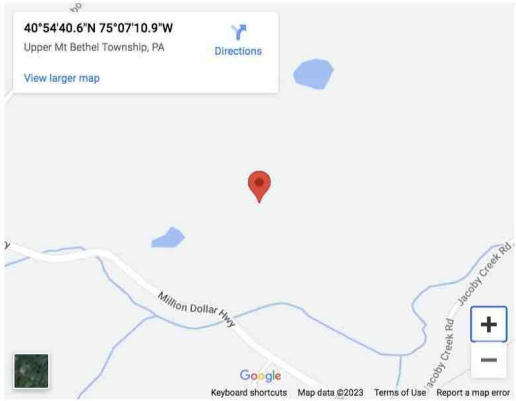
Camera settings

Make	samsung
Model	SM-G930U
Focal length	4.2 mm
Aperture	1.7
Exposure	1/2040
ISO	50
Flash	No Flash

Name	20180304_125109.jpg
File size	7.52 MB (7883594 bytes)
File type	JPEG
MIME type	image/jpeg
Image size	4032 x 3024 (12.2 megapixels)
Color space	sRGB
Created	March 04, 2018 12:51

Location

Altitude	0 m Above Sea Level
Latitude	40 deg 54' 40.62" N
Longitude	75 deg 7' 10.90" W



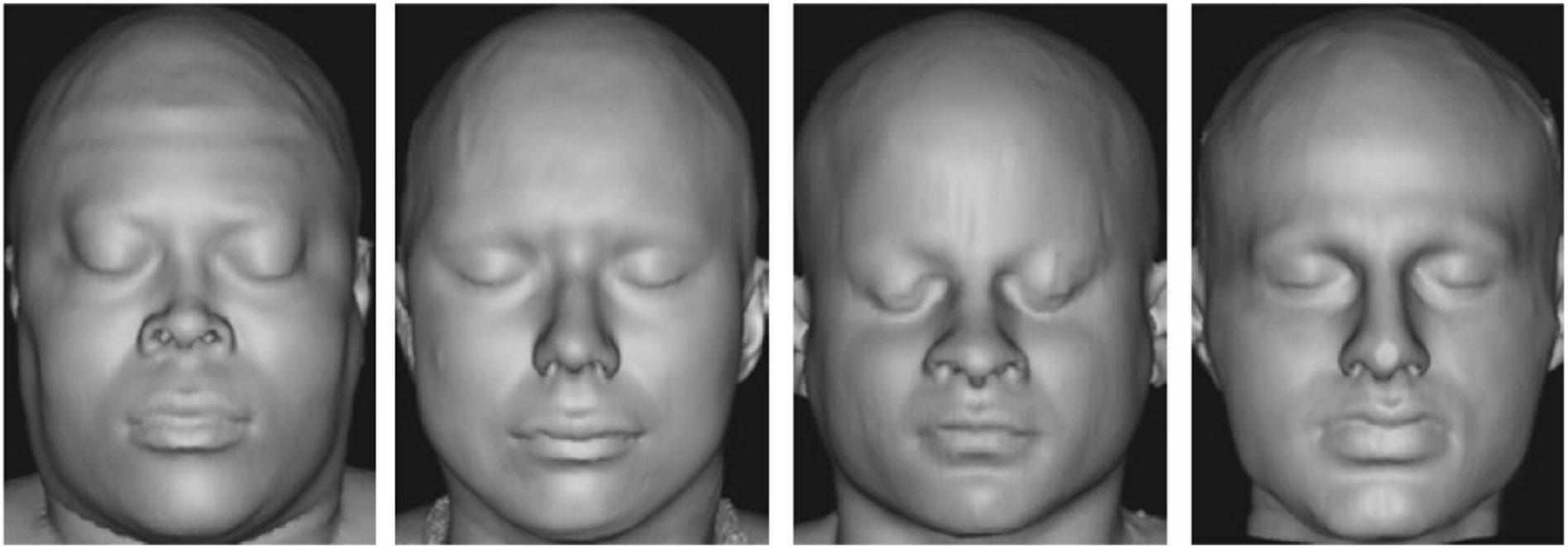
http://jimpl.com
/

Best Practice #15 - Faces (PRFI)

- *"The re-identification risk of head and neck cross-sectional images, including brain CT, MR and PET images, which may contain potentially reconstructable facial information (PRFI) that can be used by humans or facial recognition software to attempt re-identification, should be quantified with a realistic collection-specific expert statistical analysis, and if above a predetermined acceptable risk threshold, the facial features removed or modified to reduce the risk to an acceptable level, or the images should not be publicly shared"*

Best Practice #15 - Faces (PRFI)

- *"The re-identification risk of head and neck cross-sectional images, including brain CT, MR and PET images, which may contain potentially reconstructable facial information (PRFI) that can be used by humans or facial recognition software to attempt re-identification, should be quantified with a realistic collection-specific expert statistical analysis, and if above a predetermined acceptable risk threshold, the facial features removed or modified to reduce the risk to an acceptable level, or the images should not be publicly shared."*



Parks, Monson. Automated Facial Recognition of Computed Tomography-Derived Facial Images: Patient Privacy Implications.
doi:10.1007/s10278-016-9932-7

Best Practice #17 - QC

- *"A human quality control (QC) process to confirm the efficacy of the de-identification process used with respect to de-identification and preservation of utility should be used; the percentage and type of records inspected should be guided by a documented risk assessment establishing the threshold of residual risk before and after performance of the QC process. The QC process should address structured and unstructured text data elements, pixel data, geometric and bitmapped overlays, and compressed bitstream embedded metadata. The residual risk is influenced by the assessment of what is to be removed or replaced, as well as the reliability of the manner in which it is removed or replaced"*

Best Practice #17 - QC

- *"A human quality control (QC) process to confirm the efficacy of the de-identification process used with respect to de-identification and preservation of utility should be used; the percentage and type of records inspected should be guided by a documented risk assessment establishing the threshold of residual risk before and after performance of the QC process. The QC process should address structured and unstructured text data elements, pixel data, geometric and bitmapped overlays, and compressed bitstream embedded metadata. The residual risk is influenced by the assessment of what is to be removed or replaced, as well as the reliability of the manner in which it is removed or replaced."*

http://prettypetalspl.com/wp-content/uploads/2017/02/Manufacturing-and-Quality-control.jpg

Best Practice #18 - Documentation

- *"The process of de-identification used, including that performed by source sites, data coordinating centers and the entity that is responsible for the public data distribution, should be documented in detail, and that documentation, or a reference to an openly accessible source of it, published with the data collection. This documentation should include the release of the PS3.15 E.1 Application Level Confidentiality Profile used, as well as documenting any PS3.15 Confidentiality Options used."*

Best Practice #18 - Documentation

- *"The process of de-identification used, including that performed by source sites, data coordinating centers and the entity that is responsible for the public data distribution, should be documented in detail, and that documentation, or a reference to an openly accessible source of it, published with the data collection. This documentation should include the release of the PS3.15 E.1 Application Level Confidentiality Profile used, as well as documenting any PS3.15 Confidentiality Options used."*

FutUndBeidl. http://www.flickr.com/photos/61423903@N06/7382239368

Recommendation #5 - Quantify performance

- *"Further research is needed into means of quantifying the reliability of the de-identification process, whether manual or automated, such that what is intended to be removed or replaced is actually removed or replaced, and how to express this in a meaningful and understandable manner, such as by one or more "scores". This is relevant both for the consumer selecting a process, as well as comparison of different processes, such as in a competition or challenge."*

Recommendation #5 - Quantify performance

- *"Further research is needed into means of quantifying the reliability of the de-identification process, whether manual or automated, such that what is intended to be removed or replaced is actually removed or replaced, and how to express this in a meaningful and understandable manner, such as by one or more "scores". This is relevant both for the consumer selecting a process, as well as comparison of different processes, such as in a competition or challenge."*

Recommendation #8 - Actual risk of faces

- *"Further research (including thought experiments, modeling and simulations, and empirical experiments) should be performed into quantifying the actual incremental re-identification risk of potentially reconstructable facial information in head and neck cross-sectional images, to realistically assess the need for restricted access instead of public sharing, so as to balance that risk against the diminished utility of limiting access to, or de-facing such images, especially for head and neck cancer."*

Recommendation #8 - Actual risk of faces

- *"Further research (including thought experiments, modeling and simulations, and empirical experiments) should be performed into quantifying the actual incremental re-identification risk of potentially reconstructable facial information in head and neck cross-sectional images, to realistically assess the need for restricted access instead of public sharing, so as to balance that risk against the diminished utility of limiting access to, or de-facing such images, especially for head and neck cancer."*

In closing …

- The MIDI TG has made an effort to
  - summarize the state of the art
  - describe best practices
  - make recommendations for future investigation
- Recognize that there may be disagreement over some topics
  - esp. beyond the cancer community primary contributors
  - don't forget to consider unanticipated secondary uses re. preserving utility
- Residual uncertainty over key issues
  - how to conduct/perform/document risk analysis & set thresholds
  - which tools to use
  - whether to deface and how
  - …
- Feedback is still welcome
  - mistakes can be corrected in pre-print, major decisions probably not revisitable

Resources

- Report pre-print
  - http://tinyurl.com/miditgrptpre
- MIDI Task Group Wiki
  - http://tinyurl.com/NCIMIDITG
- DICOM PS3.15 Annex E
  - http://tinyurl.com/DICOM15AnnexEDeid