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

Imaging Community Call – 2022/12/05

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

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

Who is in the Task Group?

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

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."*

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."*

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 ..."*

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	(0040,4035)	N	N	X										

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

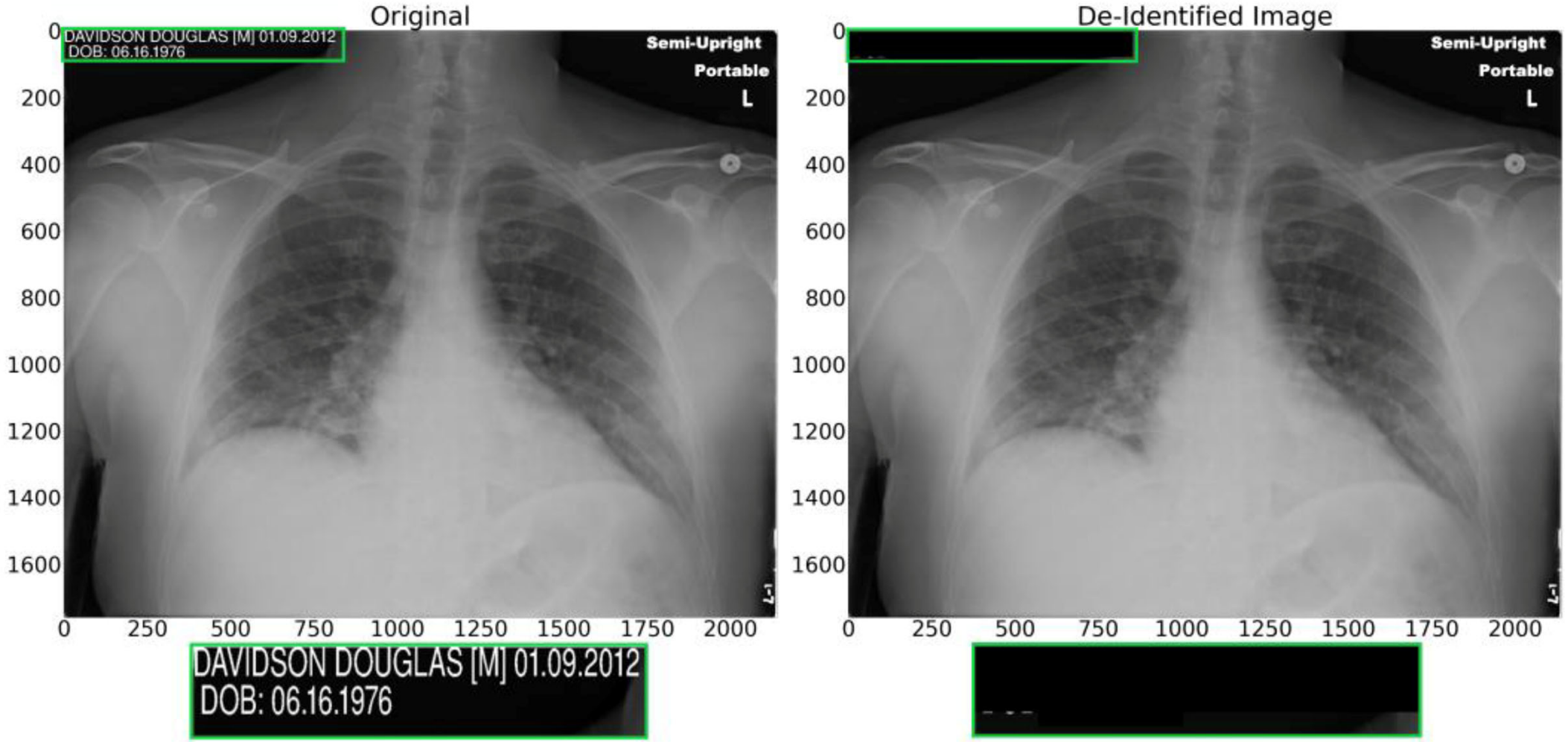
- *"For non-DICOM images, in the absence of an alternative specific reliable reference for data element retention or removal, the general principles explicit or implicit in DICOM PS3.15 E.1 should be applied, e.g., for images stored in DICOM-derived formats like Brain Imaging Data Structure (BIDS) with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

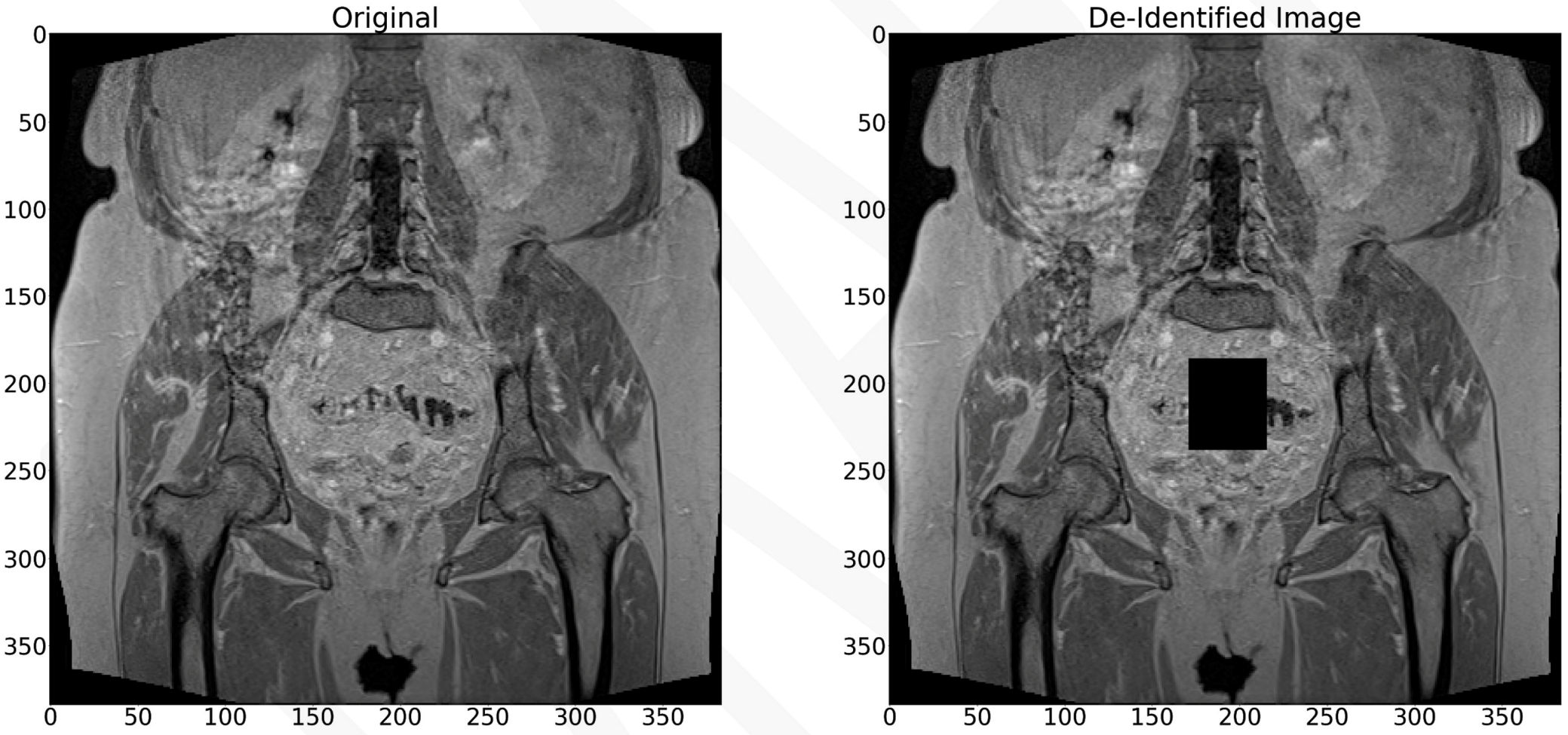
- *"Regardless of the image encoding or file format, all data elements linked to images in the collection, including those in accompanying spreadsheets or publications, which are linked by a common key (e.g., the pseudonymous subject identifier) need to be de-identified and subject to a risk analysis. 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."*



Kopchick et al. Medical image de-identification using cloud services. SPIE MI 2022. doi:10.1117/12.2608972



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

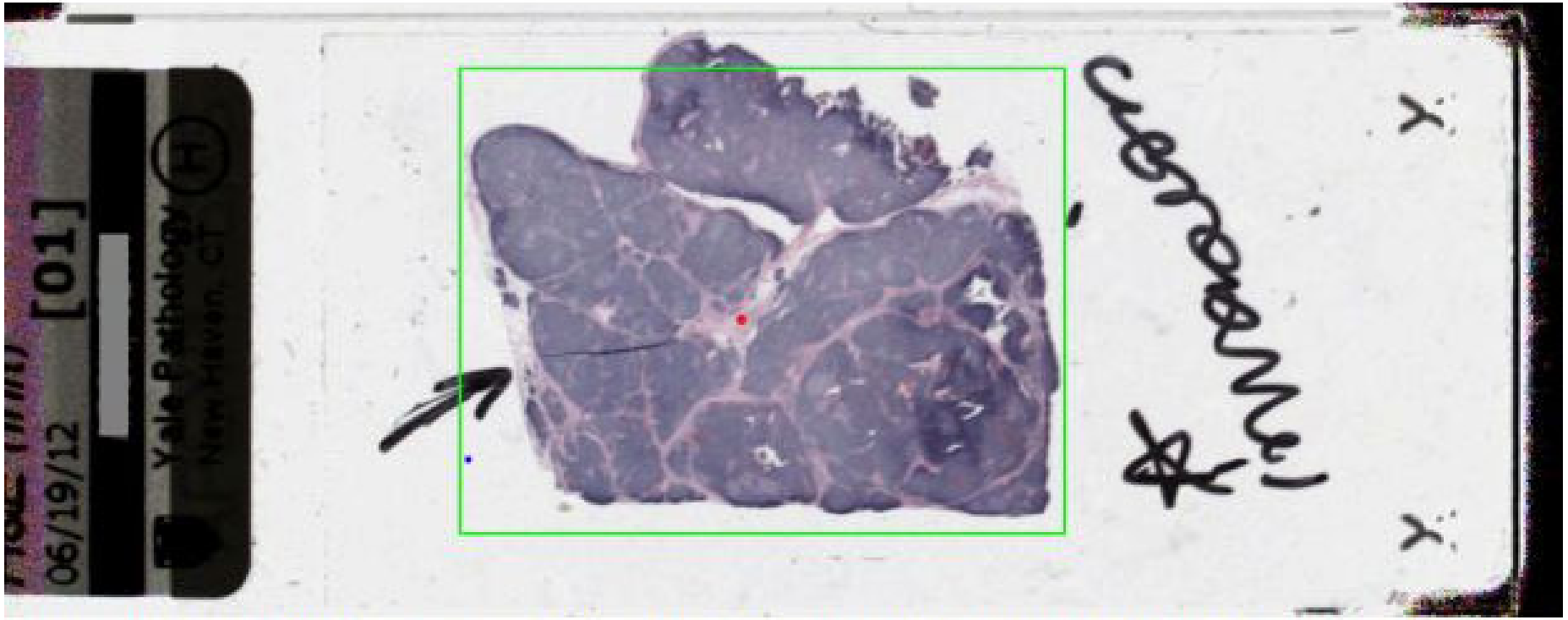


Figure 4: The macro image from TCGA file TCGA-4X-A9FB-01Z-00-DX1.211CC9AA-F721-4D16-8663-68A393223F80.svs. The left side shows part of the label which included the patient name. This has been manually redacted with a grey bar in this figure. This file was released to the public in 2016.

Gutman D. Internal report. Shared with MIDI TG 2022/04.

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 ..."*

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

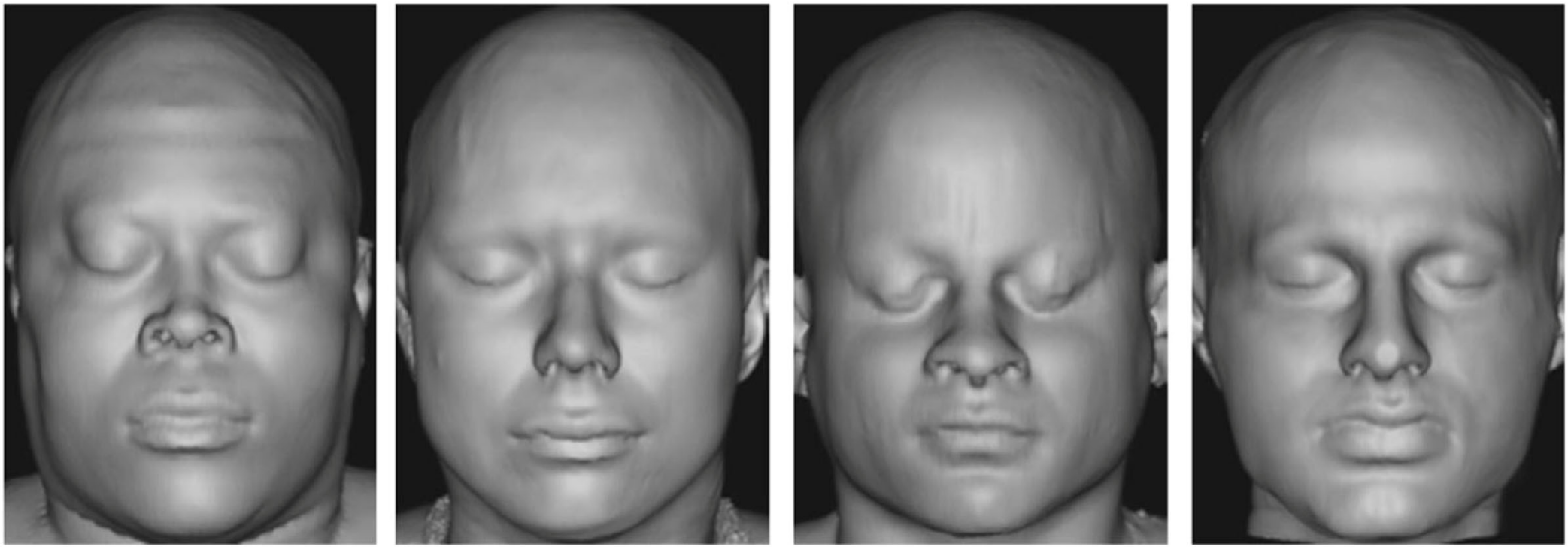
- *"Private data elements retained to preserve utility should be evaluated with respect to risk of identity leakage, either by reference to a reliable source of known safe private data elements, such as that provided in DICOM PS3.15 E.3.10, manufacturer's documentation, including DICOM Conformance Statements,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

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."*



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

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."*

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."*

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. https://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."*

Taking Dawn, Just a Taste. Single, 2018.

Please, sir, I want some more.

Oliver Twist. 1948. https://www.imdb.com/title/tt0084438/

Actions Remaining

- Continue with external review that is currently in progress
  - feedback due by 2022/12/15
  - more reviewers welcome – contact mailto:dclunie@dclunie.com
- Complete report by EOY 2022
  - distribute as pre-print (too large for academic paper)
  - prepare executive summary as academic paper
  - advertise widely
- MIDI Workshop
  - spring 2023 – exact date and F2F location in DC area TBD – will be hybrid