Medical Image Integrity Issues
International Data Security and Thoracic Screening

David Clunie, PixelMed

Disclosures

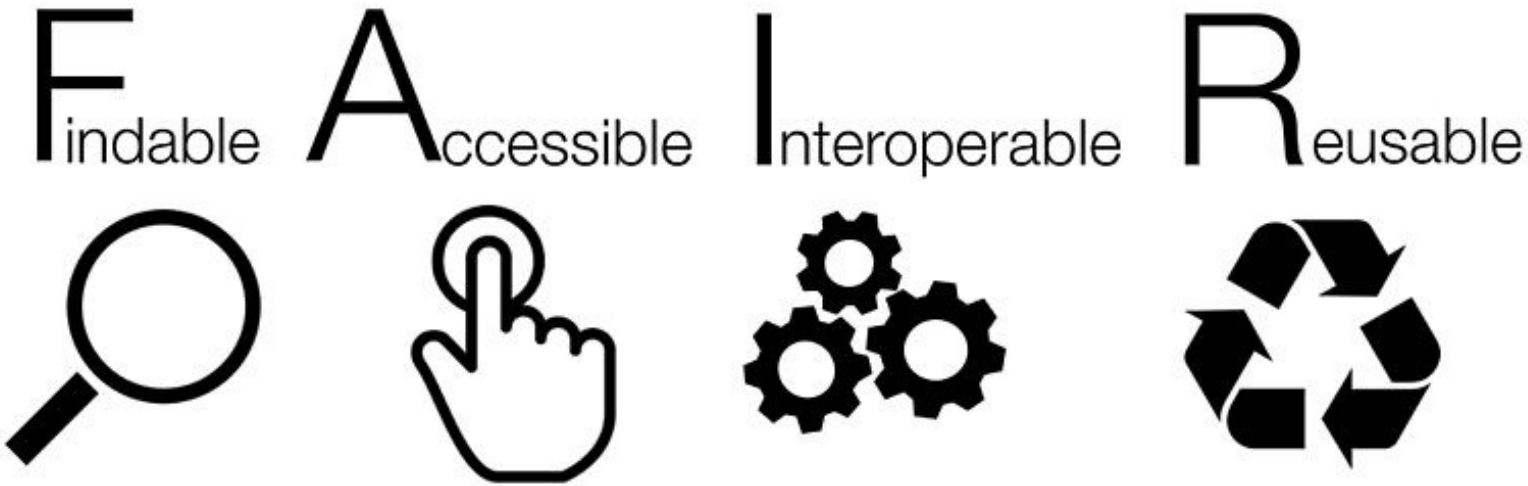
- Editor of the DICOM Standard (NEMA contract)
- NCI FNL Leidos Essex sub-contractor (SME DICOM, de-identification)
- NCI Imaging Data Commons (IDC) sub-contractor
- Consult with various equipment manufacturers regarding DICOM

Scope

- Images
  - digital – CT scans, chest X-Rays, whole slide (pathology), …
  - images are (pixel) data + metadata
- Metadata includes
  - image file “headers” – structural, identifying, descriptive
  - “associated” clinical data (spreadsheets, CSV, databases)
  - structured fields and unstructured text (including burned in)
  - annotations (patient, study, series, image, ROI, pixel, …)

- CIA “triad” – confidentiality, integrity, availability
- FAIR – findable, accessible, interoperable, and reusable
- Availability/Accessibility – standard format and protocols (DICOM)

https://www.ec-mea.com/wp-content/uploads/2019/11/CS_TriadGraphic-01-1024x900-e1574600971693.png



https://mygeohub.org/courses/fair_climate/logo


DICOM
SM
Digital Imaging and Communications in Medicine

CIA – Confidentiality

- In what context
  - active clinical data
    - needs to be identifiable
    - protected from access beyond clinical team
  - operational use as data
    - for processing by algorithm used clinically
    - for on-study patients in clinical trial
    - “transiently” de-identified/anonymi[sz]ed to those beyond clinical team
  - research (other than clinical trial)
    - internal or external, academic or commercial
    - “permanently” de-identified/anonymi[sz]ed (usually)
- What protection
  - encryption – in transit (network) and at rest (physical media, cloud)
  - de-identification/pseudonymi[sz]ation/anonymi[sz]ation

CIA – Integrity

- In what context
  - active clinical data
    - needs to be identifiable and appropriately described (w. quality control)
    - patient, staff, device, technique, anatomy, …
  - operational use as data
    - “transiently” de-identified/anonymi[sz]ed
    - preserve operational integrity – modality/anatomy/device re. algorithm
  - research (other than clinical trial)
    - “permanently” de-identified/anonymi[sz]ed (usually)
    - preserve research utility – including unanticipated secondary re-use
    - retain as much as possible at “reasonable” re-identification risk threshold
- What protection
  - reliable transport, storage, … usually deemed sufficient in practice
  - integrity checks (hashes, signatures, …) unusual, unnecessary, expensive

Overall Security Considerations

- An organizational, not just a technical problem
- People, hardware, software, …
- Many attacks by insiders (incompetent, compromised, malicious, …)
- Actual vs. perception, regulatory presumption of harm, reassurance
- Networks expanding, remote access increasing
- “Zero Trust” – same inside as outside (convenience impact)
- Organization/system is “complex” therefore vulnerable
- New is not necessarily better … just unexplored risks
- Beware of purported panaceas (“blockchain”, NFTs, …)
- Healthcare is not banking – don’t oversimplify analogy

De-identification/pseudonymi[sz]ation/anonymi[sz]ation

- Definitions & requirements vary
  - by jurisdiction, over time, …
- “Absolute” (random noise) vs “reasonable”
- Tolerance for re-identification risk
  - hard to define, quantify, establish
  - site, region, regulator, court, …
- Threat model
  - what are we trying to protect, from whom, under what constraints
- Approaches
  - “rule-based” (structured data or elements in unstructured data)
  - statistically determined (“expert determination”, “SDC”)

Computer Science > Cryptography and Security

[Submitted on 18 Mar 2023 (v1), last revised 1 Apr 2023 (this version, v2)]

Report of the Medical Image De-Identification (MIDI) Task Group -- Best Practices and Recommendations

David A. Clunie, Adam Flanders, Adam Taylor, Brad Erickson, Brian Bialecki, David Brundage, David Gutman, Fred Prior, J Anthony Seibert, John Perry, Judy Wawira Gichoya, Justin Kirby, Katherine Andriole, Luke Geneslaw, Steve Moore, TJ Fitzgerald, Wyatt Tellis, Ying Xiao, Keyvan Farahani

This report addresses the technical aspects of de-identification of medical images of human subjects and biospecimens, such that re-identification risk of ethical, moral, and legal concern is sufficiently reduced to allow unrestricted public sharing for any purpose, regardless of the jurisdiction of the source and distribution sites. All medical images, regardless of the mode of acquisition, are considered, though the primary emphasis is on those with accompanying data elements, especially those encoded in formats in which the data elements are embedded, particularly Digital Imaging and Communications in Medicine (DICOM). These images include image-like objects such as Segmentations, Parametric Maps, and Radiotherapy (RT) Dose objects. The scope also includes related non-image objects, such as RT Structure Sets, Plans and Dose Volume Histograms, Structured Reports, and Presentation States. Only de-identification of publicly released data is considered, and alternative approaches to privacy preservation, such as federated learning for artificial intelligence (AI) model development, are out of scope, as are issues of privacy leakage from AI model sharing. Only technical issues of public sharing are addressed.

Comments: 131 pages
Subjects: **Cryptography and Security (cs.CR)**; Computer Vision and Pattern Recognition (cs.CV); Image and Video Processing (eess.IV)
Cite as: arXiv:2303.10473 **[cs.CR]**
(or arXiv:2303.10473v2 **[cs.CR]** for this version)
https://doi.org/10.48550/arXiv.2303.10473

Submission history

From: David Clunie [view email]
[v1] Sat, 18 Mar 2023 19:12:38 UTC (2,006 KB)
[v2] Sat, 1 Apr 2023 16:17:40 UTC (1,586 KB)

http://tinyurl.com/miditgrptpre
http://tinyurl.com/DICOM15AnnexEDeid
http://wiki.nci.nih.gov/display/MIDI/2023+Medical+Image+De-Identification+Workshop

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 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

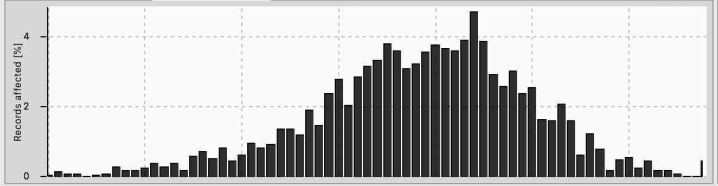
Input data | Classification performance | Quality models

	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

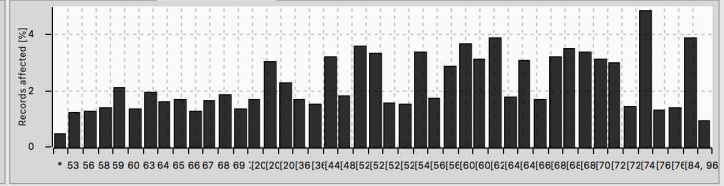
Output data | Classification performance | Quality models

	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



https://portal.imaging.datacommons.cancer.gov/explore/filters/?collection_id=nlst

IDC

https://portal.imaging.datacommons.cancer.gov/explore/

NIH NATIONAL CANCER INSTITUTE Imaging Data Commons

Explore Images | Collections | Getting Started | User Forum | News | About | Help | Sign In

Check All / Uncheck All

NLST 26408

Analysis Results

Sort by: Count Alpha

Hide analysis results with 0 cases

show 11 more Check All / Uncheck All

- TotalSegmentator-CT-Segmentations 26194
- None 1791
- nnU-Net-BPR-annotations 571
- QIN-LungCT-Seg 0
- RIDER-LungCT-Seg 0

Search Configuration

Hide attribute values with 0 cases

ORIGINAL DERIVED RELATED

- Primary Site Location
- Cancer Type
- Body Part Examined

Modality

Find studies with ANY ALL selected modalities

Sort by: Count Alpha

show 21 more Check All / Uncheck All

- Computed Tomography 26254
- SR Document 26194
- Segmentation 26194
- Slide Microscopy 449

Charts

Collections

Showing 1 to 1 of 1 entries Show 10 entries Previous 1 Next Search:

Collection Name	Total # of Cases	# of Cases(this cohort)
NLST	26408	26408

Selected Cases

Showing 0 to 0 of 0 entries Show 10 entries Page Go Previous Next Find by Case ID: Go

Collection Name	Case ID	Total # of Studies	Total # of Series
No data available in table			

Selected Studies

Showing 0 to 0 of 0 entries Show 10 entries Page Go Previous Next Find by Study Instance UID: Go

Case ID	Study Instance UID	Study Date	Study Description	# of Series	View
No data available in table					

Selected Series

Additional concerns in an AI-enabled world

- Greater demand for data access – training, inference, QC, regulator
- Data has a different “value” now
- Reproducibility, generalizability (subjects, exposure, protocols, machines, …)
- Models are data too – may also be privacy leak
- In-house AI means in-house foreign code – security risk
- Greater IT complexity to support it – inherently less secure
- Beyond in-house review capability/expertise
- AI in the cloud (OTS providers, no on-premise HPC, network cost, but trust?)
- Provider audit, security penetration tests need extending to imaging, AI, research, de-identification, …
- AI as a de-identification tool, and as a re-identification tool

https://i2.wp.com/www.geekinlibrariansclothing.com/wp-content/uploads/2016/01/conspiracy-quotes-x-Files-Trust-No-One.jpg