S5-DLL02 – RSNA Deep Learning Lab

DICOM De-Identification Using ChatGPT

Errol Colak, MD, FRCPC

Department of Medical Imaging, St. Michael's Hospital and University of Toronto

TORONTO RADIOLOGY

ST. MICHAEL'S UNITY HEALTH TORONTO

No Disclosures

Objectives

- Private health information
- DICOM
- Data de-identification
- Lessons learned
- Best practices

Why do we care about data de-identification?

What is Protected Data?

- General Data Protection Regulation (GDPR)
  - Any data that relates to, or can lead to the identification of a living person
- Health Insurance Portability and Accountability Act (HIPAA)
  - Any information about health status, care, or payment that is created or collected by a HIPAA covered entity, that can be linked to a specific individual

Protected Health Information (PHI)

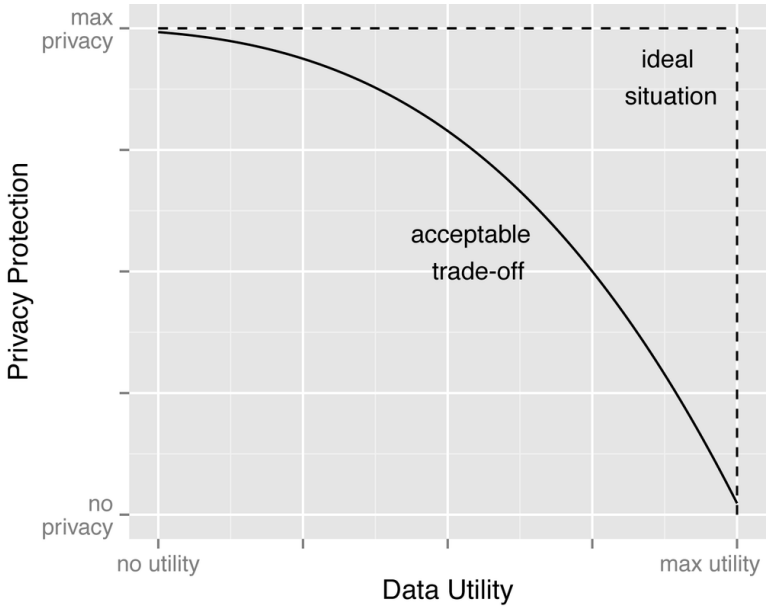
- What is it?
- Why does protecting it matter?
  - Legal and regulatory compliance
  - Ethical considerations
  - Building patient trust
  - Organizational reputation

Data De-identification

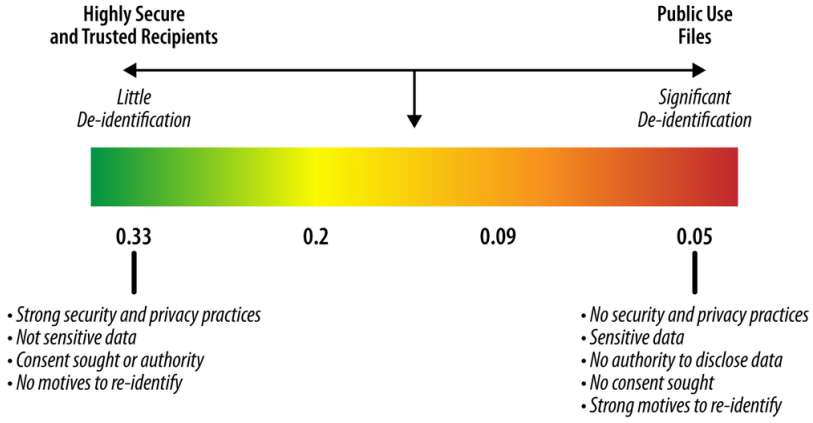
- Process by which information cannot be associated with a particular individual
- Facilitate research, education, & data sharing
- Protecting patient confidentiality
- Compliance with regulations
- Minimizing the risk of data breaches
- Secondary data usage

Data De-identification

- Key principles:
  - Removal of identifiable information
  - Preservation of research and clinical value
- Challenges:
  - Balancing anonymity and utility
  - Image quality and integrity
  - Large datasets in medical imaging



Emam, K. E., & Arbuckle, L. (2014). Anonymizing Health Data. O'Reilly Media.



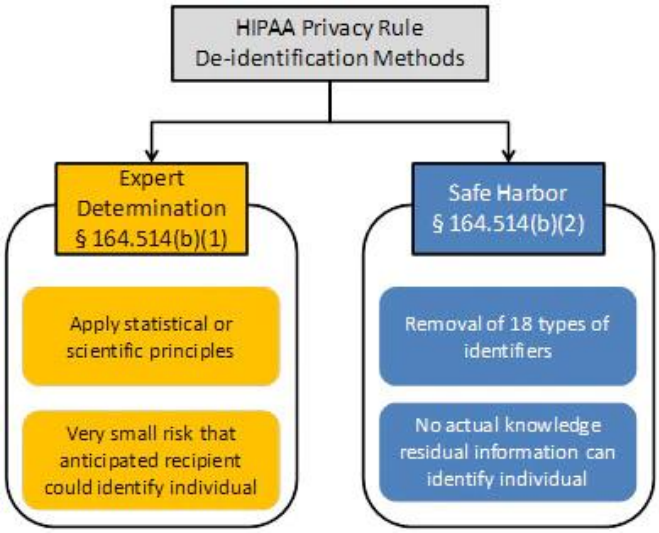
Emam, K. E., & Arbuckle, L. (2014). Anonymizing Health Data. O'Reilly Media.

Terminology

- De-identification, anonymization, pseudonymization
- De-identification:
  - Protect privacy by dissociating data from specific individuals
  - Allows for data use and analysis for legitimate purposes
- Anonymization:
  - Goes a step further than de-identification
  - Ensures that the data cannot be linked back to specific individuals through any means, direct or indirect

Approaches to Protecting PHI

- Complete removal of all PHI
- Replacing specific identifying elements with general descriptors or placeholders
- Generalization
- Pixelization or masking
- Encryption
- Noise Addition



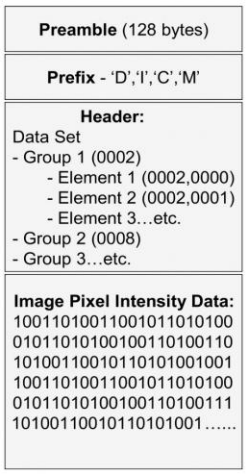
https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

DICOM

- Digital Imaging and Communications in Medicine
- Comprehensive framework
  - Data structure
  - Data encoding
  - Metadata
  - Communication protocol
  - Image display and processing
  - Workflow and interoperability

DICOM Files

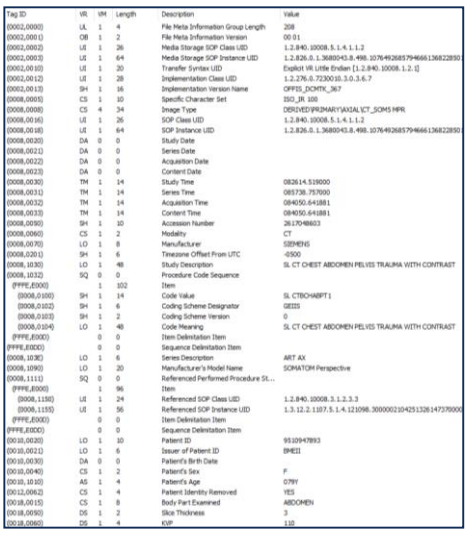
- Each file is designed to be standalone
- Two main components:
  - Header (metadata)
  - Pixel data



Varma DR. Indian J. Radiol. Imaging. 2012 Jan;22(01):4-13.

DICOM Tags

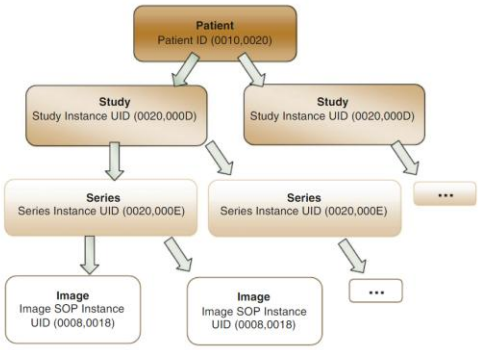
- Public
- Private
  - May contain PHI



Tag ID	VR	VM	Length	Description	Value
(0002,0000)	UL	1	4	File Meta Information Group Length	208
(0002,0001)	OB	1	2	File Meta Information Version	00 01
(0002,0002)	UI	1	26	Media Storage SOP Class UID	1.2.840.10008.5.1.4.1.1.2
(0002,0003)	UI	1	64	Media Storage SOP Instance UID	1.2.826.0.1.3680043.8.498.10764926857946661368228501
(0002,0010)	UI	1	20	Transfer Syntax UID	Explicit VR Little Endian [1.2.840.10008.1.2.1]
(0002,0012)	UI	1	28	Implementation Class UID	1.2.276.0.7230010.3.0.3.6.7
(0002,0013)	SH	1	16	Implementation Version Name	OFFIS_DCMTK_367
(0008,0005)	CS	1	10	Specific Character Set	ISO_IR 100
(0008,0008)	CS	4	34	Image Type	DERIVED\PRIMARY\AXIAL\CT_SOMS MPR
(0008,0016)	UI	1	26	SOP Class UID	1.2.840.10008.5.1.4.1.1.2
(0008,0018)	UI	1	64	SOP Instance UID	1.2.826.0.1.3680043.8.498.10764926857946661368228501
(0008,0020)	DA	0	0	Study Date	
(0008,0021)	DA	0	0	Series Date	
(0008,0022)	DA	0	0	Acquisition Date	
(0008,0023)	DA	0	0	Content Date	
(0008,0030)	TM	1	14	Study Time	082614.519000
(0008,0031)	TM	1	14	Series Time	085738.757000
(0008,0032)	TM	1	14	Acquisition Time	084050.641881
(0008,0033)	TM	1	14	Content Time	084050.641881
(0008,0090)	SH	1	10	Accession Number	2817048603
(0008,0060)	CS	1	2	Modality	CT
(0008,0070)	LO	1	8	Manufacturer	SIEMENS
(0008,0201)	SH	1	6	Timezone Offset From UTC	-0500
(0008,1030)	LO	1	48	Study Description	SL CT CHEST ABDOMEN PELVIS TRAUMA WITH CONTRAST
(0008,1032)	SQ	0	0	Procedure Code Sequence	
(FFFE,E000)		1	102	Item	
(0008,0100)	SH	1	14	Code Value	SL CTBCHABPT 1
(0008,0102)	SH	1	6	Coding Scheme Designator	GEIIS
(0008,0103)	SH	1	2	Coding Scheme Version	0
(0008,0104)	LO	1	48	Code Meaning	SL CT CHEST ABDOMEN PELVIS TRAUMA WITH CONTRAST
(FFFE,E00D)		0	0	Item Delimitation Item	
(FFFE,E0DD)		0	0	Sequence Delimitation Item	
(0008,103E)	LO	1	6	Series Description	ART AX
(0008,1090)	LO	1	20	Manufacturer's Model Name	SOMATOM Perspective
(0008,1111)	SQ	0	0	Referenced Performed Procedure St...	
(FFFE,E000)		1	96	Item	
(0008,1150)	UI	1	24	Referenced SOP Class UID	1.2.840.10008.3.1.2.3.3
(0008,1155)	UI	1	56	Referenced SOP Instance UID	1.3.12.2.1107.5.1.4.121098.30000021042513261473700001
(FFFE,E00D)		0	0	Item Delimitation Item	
(FFFE,E0DD)		0	0	Sequence Delimitation Item	
(0010,0020)	LO	1	10	Patient ID	9510947893
(0010,0021)	LO	1	6	Issuer of Patient ID	BMEII
(0010,0030)	DA	0	0	Patient's Birth Date	
(0010,0040)	CS	1	2	Patient's Sex	F
(0010,1010)	AS	1	4	Patient's Age	079Y
(0012,0062)	CS	1	4	Patient Identity Removed	YES
(0018,0015)	CS	1	8	Body Part Examined	ABDOMEN
(0018,0050)	DS	1	2	Slice Thickness	3
(0018,0060)	DS	1	4	KVP	110

DICOM Hierarchy

- Patient
  - Study
    - Series
      - Image



Pianykh, O.S. (2012). Parlez-vous DICOM?. In: Digital Imaging and Communications in Medicine (DICOM). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10850-1_5

De-Identification of DICOM Images

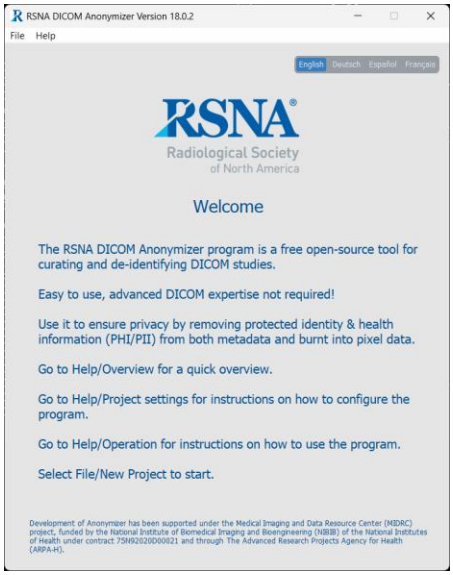
- Patient level de-identification
  - Removing or replacing patient-specific identifiers (name, ID, etc.) with pseudonyms or anonymized values
  - Anonymizing demographics (e.g. age, sex) by altering values within an acceptable range.
- Study and series level de-identification
  - Remove study and series descriptions that may reveal sensitive information
  - Remove/modify study dates and times
  - Modify/remove acquisition parameters that could indirectly identify a patient
- Image level de-identification
  - Remove annotations/overlays that contain PHI or identifiable information
  - Blurring or obfuscating regions containing patient-specific features
  - Adjusting pixel values to remove identifiable patterns
  - Ensure modifications maintain overall diagnostic image quality

Process Overview

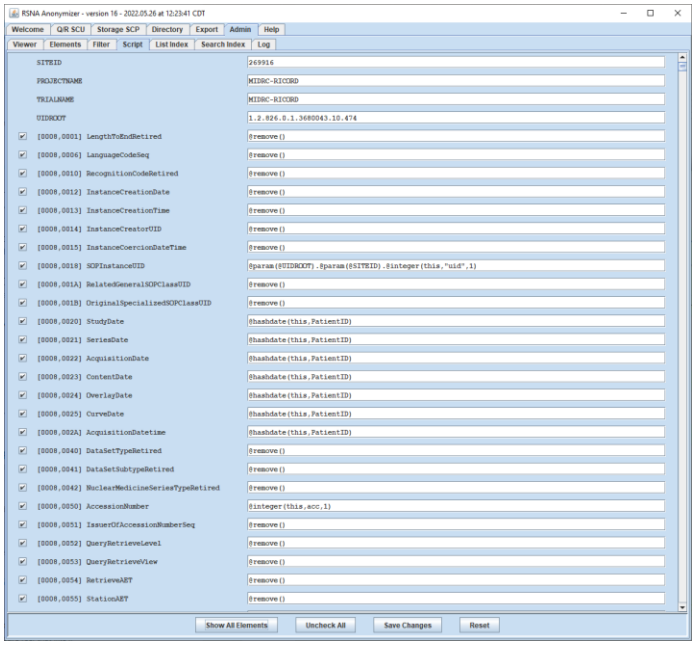
- Understand DICOM file
- Review privacy regulations and local governance
  - Identify sensitive dataelements
  - Define de-identification policies and methods
- Establish a secure environment
- Pre-de-identification data validation
- Execute de-identification process
- Post-de-identification data validation
- Maintain data security and compliance
- Document, periodic audits, and quality assurance
- Stay Informed and updated

De-Identification Tools

- RSNA Anonymizer
- RSNA Clinical Trial Processor (CTP)
- DicomCleaner
- XNAT
- Orthanc
- Python



https://github.com/RSNA/Anonymizer



Whitelisting

De-Identified DICOM file

1. Copy image matrix
2. Copy over metadata tags specified in a safe list

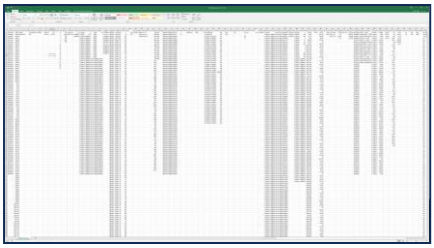
Whitelisted DICOM file

Quality Assurance and Validation

- Develop a robust QA process
- Documentation and standard operating procedures
- Manual review and sample audits
- Automated validation tools
- Metadata dump

Metadata Dump

AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
KVP	Manufact	Manufact	Modality	NumberO	PatientAg	PatientID	PatientNa	PatientPo	PatientSe
140	GE MEDIC	Optima C	CT	729	050Y	269916-00	269916-00	HFS	F
120		Revolution	CT	1266	080Y	269916-00	269916-00	FFS	M
		LightSpeed	VCT	2345	060Y	269916-00	269916-000655		
				713	120Y	269916-00	269916-000674		
				1594	055Y	269916-00	269916-000682		
				673	030Y	269916-00	269916-000686		
				536	075Y	269916-00	269916-000699		
				1916	040Y	269916-00	269916-000707		
				720	035Y	269916-00	269916-000782		
				1201	045Y	269916-00	269916-000792		
				752	065Y	269916-00	269916-000799		
				1572	070Y	269916-00	269916-000804		
				710	090Y	269916-00	269916-000805		
				669	025Y	269916-00	269916-000820		
				734	085Y	269916-00	269916-000840		
				1837		269916-00	269916-000846		
				1389		269916-00	269916-000879		
				1255		269916-00	269916-000882		
				1729		269916-00	269916-000937		
				1854		269916-00	269916-000987		
				1853		269916-00	269916-001004		
				4179		269916-00	269916-001011		
				834		269916-00	269916-001015		
				644		269916-00	269916-001070		



Best Practices and Considerations

- Documentation of the de-identification process
- Collaboration with IT and compliance departments
- Use validated tools
- Training and education
- Regular updates and compliance audits

Lessons Learned

- Carefully estimate resources required
- Dates require special attention
- Do not neglect documentation
- Do not rely on DICOM metadata to indicate burned-in PHI
- Do not overdo de-identification

Take Away Points

- De-identification is a multi-layered strategy
- Be familiar with DICOM file format
- Use validated de-identification tools
- Seek advice from experienced individuals

Errol.Colak@UnityHealth.to