TCIA De-identification and Curation - Public

TCIA De-identification and Image Curation

30 June, 2015

Overview - Following industry best-practices, TCIA uses a standards-based approach to de-identification of DICOM images to insure that images are free of protected health information (PHI). The TCIA de-identification process ensures that the HIPAA de-identification standard is met by following the Safe Harbor Method as defined in section 164.514(b)(2) of the HIPPA Privacy Rule. The standard for de-identification of DICOM objects is defined by the DICOM Standard PS 3.15-2011 Digital Imaging and Communications in Medicine (DICOM), Part 15: Security and System Management Profiles (medical.nema.org/Dicom/2011/11_15pu.pdf). At the submitting site, a DICOM PS 3.15 compliant script removes or modifies DICOM tags deemed to be unsafe (See table 1 for a complete listing). TCIA incorporates the “Basic Application Confidentiality Profile” which is amended by inclusion of the following profile options: Clean Pixel Data Option, Clean Descriptors Option, Retain Longitudinal With Modified Dates Option, Retain Patient Characteristics Option, Retain Device Identity Option, and Retain Safe Private Option. The de-identification rules applied to each object are recorded by TCIA in the DICOM sequence Method Code Sequence [0012,0063] by entering the Code Value, Coding Scheme Designator, and Code Meaning for each profile and option that were applied to the DICOM object during de-identification. The DICOM standard for de-identification of objects defines a minimum set of elements to de-identify to be in compliance with the standard. It is up to the user doing the de-identification to insure that PHI is removed or cleaned according to the laws and practices in place at the time de-identification occurs.

Base level de-identification - The Basic Application Confidentiality Profile requires that Patient Name and Patient ID are either blanked or modified. TCIA incorporates an ID mapping between the original Patient ID and the ID that the images will have within TCIA. The mapping table is created at the image submitting site, the mapping performed prior to the images leaving the sites host computer, and TCIA never sees the original Patient ID. The remapped Patient ID is also mapped to the Patient Name field. This is done for the case where a DICOM viewer or application being used by the TCIA user that downloaded the data would require a Patient Name to be present. To show that the Patient Identity has been removed, the term “YES” is written into DICOM tag 00120062 “PatientIdentityRemoved”.

In general, the Basic Application Profile specifies removal or modification of any tag that by definition would contain PHI that could be used either alone or together with other information to uniquely identify a subject. Removal of detailed geographic information, dates, exam identifiers, patient demographics, free text entry fields, vendor private tags, etc. are all done to minimize the possibility of being able to uniquely identify an individual. The options to the DICOM de-identification standard allows for retention of information to help make the data scientifically valuable, but as more options are added the chance of PHI is increased and a rigorous de-identification process must be followed.

Exam Identifiers - DICOM makes extensive use of universal identifiers (UID) that could be used to identify a subject if a user had access to the PACS system at the institution where the images originated. The Basic Application Confidentiality Profile requires that all UIDs be removed or modified. TCIA uses its own root UID, appends an 8 digit string in the form of xxxx.yyyy (where xxxx is related to the collection and yyyy is related to a submitting site) and then appends a hashed value of the original UID. UIDs have no special meaning other than serving as unique identifiers and the only reason TCIA adds the 8 digit string is to minimize the possibility of two images being assigned the same UID as images come from many different sites. This technique insures that images stay associated with the appropriate series, study, and subject as well as ensuring that referenced images between secondary capture images, structured reports, PET/CT, etc. are still valid references to images within TCIA. Any image resubmitted to TCIA will have the same UID to avoid the same image appearing twice with a different identifier. Original accession numbers are hashed with a 16 bit string to prevent linking of DICOM objects back to the submitting site.

Dates - The Retain Longitudinal With Modified Dates Option allows dates to be retained as long as they are modified from the original date. Date and Date-Time fields in TCIA DICOM image headers have been offset based on a random number, but the longitudinal relationship between dates is maintained. Therefore, a researcher won't know the precise date the scan occurred, but if a follow up scan was performed 120 days later, that same 120 day difference between scans of a subject will exist in the TCIA images. Dates that occur in DICOM tags other than Date or Date-Time fields are removed. An example of this would be a date entered into the Series Description field. If the date is associated with a library for Code Meaning then that date is preserved as the date would be required to look up the meaning in the correct version of the library. To show that the dates have been modified, the term "MODIFIED" is written into DICOM tag 00280303 "LongitudinalTemporalInformationModified".

Patient Demographics – The keep Patient Characteristics Option allows keeping some patient demographics for research purposes. The allowed fields are Patient's Sex, Patient's Age, Patient's Size, Patient's Weight, Ethnic Group, Smoking Status, and Pregnancy Status. If a subject is over 90 years of age, then the age must be listed as 90+. Allergies, Patient State (this is not where they live, rather their condition), Pre-Medication, and Special Needs are defined by the DICOM standard as "clean" and are kept by TCIA and examined for PHI along with all tags during curation. Other patient demographics such as birthdate, address, religious affiliations, etc. are removed or emptied.

The names of health care providers including staff, hospital name, assigned IDs etc. are removed from the DICOM objects in cases where there is enough detail to identify an individual or facility where the scan was done.

Free Text - The Clean Descriptors Option allows for DICOM tags where free text could be entered by a technician to be kept. The following tags fall under that option and are all kept, inspected, and cleaned of PHI by TCIA during the curation process: Allergies, Patient State, Study Description, Series Description, Admitting Diagnoses Description, Admitting Diagnoses Code Sequence, Derivation Description, Identifying Comments, Medical Alerts, Occupation, Additional Patient's History, Patient Comments, Contrast Bolus Agent, Protocol Name, Acquisition Device Processing Description, Acquisition Comments, Acquisition Protocol Description, Contribution Description, Image Comments, Frame Comments, Reason for Study, Requested Procedure Description, Requested Contrast Agent, Study Comments, Discharge Diagnosis Description, Service Episode Description, Visit Comments, Scheduled Procedure Step Description, Performed Procedure Step Description, Comments on Performed Procedure Step, Requested Procedure Comments, Reason for Imaging Service Request, Imaging Service Request Comments, Interpretation Text, Interpretation Diagnosis Description, Impressions, and Results Comments. The TCIA de-identification script run at the submitting sites removes the field "Request Attributes Sequence" as that tag typically contains PHI and provides no scientific value. Many of these fields contain information valuable to research and are important to retain. For images that are submitted with missing Series Descriptions, TCIA will add text to Series Descriptions to help researchers during TCIA image searches. When a missing series description is encountered, TCIA staff will use the following approach: Enter "LOCALIZER" if the ImageType contains the word localizer; Enter "Contrast" and then append the value contained in Contrast Bolus Agent if a value is present; if Contrast Bolus Agent is missing or empty other tags will be examined to see if a series was scanned with contrast (The Image Comments field is often used by sites to denote contrast); if the Image is an MR then TCIA will map the Scanning Sequence parameters into the Series Description; if none of those conditions apply then TCIA will map Scan Options or simply enter "none" into the Series Description field.

Devices - The Retain Device Identity Option of the DICOM de-identification standard allows for the retention of information related to the scanner used. The option allows for the following relevant tags to be retained: Station Name, Device Serial Number, Device UID, Plate ID, Generator ID, Cassette ID, Gantry ID, Detector ID, Scheduled Study Location, Scheduled Study Location AE Title, Scheduled Station AE Title, Scheduled Station Name, Scheduled Procedure Step Location, Performed Station AE Title, Performed Station Name, Performed Station Name Code Sequence, Scheduled Station Name Code Sequence, Scheduled Station Geographic Location Code Sequence, and Performed Station Geographic Location Code Sequence. TCIA removes Station Name as part of its de-identification process as Station Name often contains information related to the site where the scan occurred. The other tags listed above are retained if they are found to be free of PHI after TCIA curation of the submitted DICOM objects.

Private Tags - When a submitting site sends DICOM data to TCIA all private tags are retained and then de-identified by TCIA during curation of the data according to the Retain Safe Private Option. The Retain Safe Private Option allows for the retention of DICOM tags stored in the private fields. These fields are extensively used by DICOM vendors to store information about the scans. To claim conformance to the DICOM standard the vendors must publish a DICOM conformance statement that defines the standard and private tags that are used by their particular equipment. These conformance statements are typically made available on the vendors website for download. Unfortunately, there are cases where vendors do not make the conformance statement for a piece of equipment publicly available or do not adequately define what is stored in the private tags. In TCIA the Private DICOM elements are de-identified according to the rules contained in a de-identification knowledge base maintained by the TCIA team at the University of Arkansas for Medical Sciences https://queries.cancerimagingarchive.net/PrivateElementKnowledgeBase/faces/index.xhtml. This knowledge base defines rules for de-identification of private tags based on a vendor's conformance statement for each scanner and software version. The manufacturer, manufacturer model, modality, and software version are extracted from each series submitted. The TCIA de-identification knowledge base is checked for a conformance statement matching these data. If not found, TCIA locates the conformance statement and adds it to the knowledge base. TCIA will remove any private tags from the images that are not specified in the conformance statement or are defined as containing a form of PHI such as name, SSN, etc. All date and datetime private tags that are retained are offset using the same offset as applied to the standard tags for the image. All private tags containing UIDs are assigned a TCIA root and appended with a hashed value as done with the standard tags. This ensures all references to other images contained within TCIA are maintained. A manual inspection of all private tags is performed using tagSniffer reports and any PHI that may be found is removed, emptied, date offset, or hashed as appropriate.

Body Part Examined - When images are made public, a single body part examined, corresponding to the cancer of interest, is assigned to all images. If the collection consists of sarcoma images (or any other cancer affecting multiple organs within the image collection), there may be multiple body parts assigned, though only one to any series. In phantom collections, body part examined is simply labeled "PHANTOM".

All Tags - The TCIA de-identification process ensures that every DICOM tag of every DICOM object is free of the 18 forms of PHI as currently defined by the Safe Harbor Method. At the submitting site, a DICOM PS 3.15 compliant script removes or modifies DIOCM tags deemed to be unsafe (See table 1 for a complete listing). At TCIA, a software routine known as tagSniffer extracts every unique value found within a collection being curated and prints them to a report. This report is examined by curators and any actions necessary to remove PHI is applied when moving the images from the Intake server to the Public Server. Every DICOM image is inspected by curators for burned in PHI. Once the images reach the Public Server, the tags are inspected by two curators for PHI using new tagSniffer reports. Images are spot checked for any burned in PHI.

The following table details the de-identification performed at the submitting site by way of a TCIA supplied de-identification script.

Table 1

Tag	Name	Action
00080050	AccessionNumber	hash
00184000	AcquisitionComments	keep
00400555	AcquisitionContextSeq	remove
00080022	AcquisitionDate	incrementdate
0008002a	AcquisitionDatetime	incrementdate
00181400	AcquisitionDeviceProcessingDescription	keep

00189424	AcquisitionProtocolDescription	keep
00080032	AcquisitionTime	keep
00404035	ActualHumanPerformersSequence	remove
001021b0	AdditionalPatientHistory	keep
00380010	AdmissionID	remove
00380020	AdmittingDate	incrementdate
00081084	AdmittingDiagnosesCodeSeq	keep
00081080	AdmittingDiagnosesDescription	keep
00380021	AdmittingTime	keep
00102110	Allergies	keep
40000010	Arbitrary	remove
0040a078	AuthorObserverSequence	remove
00130010	BlockOwner	CTP
00180015	BodyPartExamined	BODYPART
00101081	BranchOfService	remove
00280301	BurnedInAnnotation	keep
00181007	CassetteID	keep
00400280	CommentsOnPPS	keep
00209161	ConcatenationUID	hashuid
00403001	ConfidentialityPatientData	remove
00700086	ContentCreatorsIdCodeSeq	remove
00700084	ContentCreatorsName	empty
00080023	ContentDate	incrementdate
0040a730	ContentSeq	remove
00080033	ContentTime	keep
0008010d	ContextGroupExtensionCreatorUID	hashuid
00180010	ContrastBolusAgent	keep
0018a003	ContributionDescription	keep
00102150	CountryOfResidence	remove

00089123	CreatorVersionUID	hashuid
00380300	CurrentPatientLocation	remove
00080025	CurveDate	incrementdate
Group	curves	remove
00080035	CurveTime	keep
0040a07c	CustodialOrganizationSeq	remove
fffcfffc	DataSetTrailingPadding	remove
00181200	DateofLastCalibration	incrementdate
0018700c	DateofLastDetectorCalibration	incrementdate
00181012	DateOfSecondaryCapture	incrementdate
00120063	DeIdentificationMethod	{Per DICOM PS 3.15 AnnexE. Details in 0012,0064}
00120064	DeIdentificationMethodCodeSequence	113100/113101/113105/113107/113108/113109/113111
00082111	DerivationDescription	keep
0018700a	DetectorID	keep
00181000	DeviceSerialNumber	keep
00181002	DeviceUID	keep
fffafffa	DigitalSignaturesSeq	remove
04000100	DigitalSignatureUID	remove
00209164	DimensionOrganizationUID	hashuid
00380040	DischargeDiagnosisDescription	keep
4008011a	DistributionAddress	remove
40080119	DistributionName	remove
300a0013	DoseReferenceUID	hashuid
00102160	EthnicGroup	keep
00080058	FailedSOPInstanceUIDList	hashuid
0070031a	FiducialUID	hashuid
00402017	FillerOrderNumber	empty
00209158	FrameComments	keep
00200052	FrameOfReferenceUID	hashuid

00181008	GantryID	keep
00181005	GeneratorID	keep
00700001	GraphicAnnotationSequence	remove
00404037	HumanPerformersName	remove
00404036	HumanPerformersOrganization	remove
00880200	IconImageSequence	remove
00084000	IdentifyingComments	keep
00204000	ImageComments	keep
00284000	ImagePresentationComments	remove
00402400	ImagingServiceRequestComments	keep
40080300	Impressions	keep
00080012	InstanceCreationDate	incrementdate
00080014	InstanceCreatorUID	hashuid
00080081	InstitutionAddress	remove
00081040	InstitutionalDepartmentName	remove
00080082	InstitutionCodeSequence	remove
00080080	InstitutionName	remove
00101050	InsurancePlanIdentification	remove
00401011	IntendedRecipientsOfResultsIDSequence	remove
40080111	InterpretationApproverSequence	remove
4008010c	InterpretationAuthor	remove
40080115	InterpretationDiagnosisDescription	keep
40080202	InterpretationIdIssuer	remove
40080102	InterpretationRecorder	remove
4008010b	InterpretationText	keep
4008010a	InterpretationTranscriber	remove
00083010	IrradiationEventUID	hashuid
00380011	IssuerOfAdmissionID	remove
00100021	IssuerOfPatientID	remove

00380061	IssuerOfServiceEpisodeId	remove
00281214	LargePaletteColorLUTUid	hashuid
001021d0	LastMenstrualDate	incrementdate
00280303	LongitudinalTemporalInformationModified	MODIFIED
04000404	MAC	remove
00080070	Manufacturer	keep
00081090	ManufacturerModelName	keep
00102000	MedicalAlerts	keep
00101090	MedicalRecordLocator	remove
00101080	MilitaryRank	remove
04000550	ModifiedAttributesSequence	remove
00203406	ModifiedImageDescription	remove
00203401	ModifyingDeviceID	remove
00203404	ModifyingDeviceManufacturer	remove
00081060	NameOfPhysicianReadingStudy	remove
00401010	NamesOfIntendedRecipientsOfResults	remove
00102180	Occupation	keep
00081070	OperatorName	remove
00081072	OperatorsIdentificationSeq	remove
00402010	OrderCallbackPhoneNumber	remove
00402008	OrderEnteredBy	remove
00402009	OrderEntererLocation	remove
04000561	OriginalAttributesSequence	remove
00101000	OtherPatientIDs	remove
00101002	OtherPatientIDsSeq	remove
00101001	OtherPatientNames	remove
00080024	OverlayDate	incrementdate
Group	overlays	remove
00080034	OverlayTime	keep

00281199	PaletteColorLUTUID	hashuid
0040a07a	ParticipantSequence	remove
00101040	PatientAddress	remove
00101010	PatientAge	keep
00100030	PatientBirthDate	empty
00101005	PatientBirthName	remove
00100032	PatientBirthTime	remove
00104000	PatientComments	keep
00100020	PatientID	Re-Mapped
00120062	PatientIdentityRemoved	YES
00380400	PatientInstitutionResidence	remove
00100050	PatientInsurancePlanCodeSeq	remove
00101060	PatientMotherBirthName	remove
00100010	PatientName	Re-Mapped
00102154	PatientPhoneNumbers	remove
00100101	PatientPrimaryLanguageCodeSeq	remove
00100102	PatientPrimaryLanguageModifierCodeSeq	remove
001021f0	PatientReligiousPreference	remove
00100040	PatientSex	keep
00102203	PatientSexNeutered	keep
00101020	PatientSize	keep
00380500	PatientState	keep
00401004	PatientTransportArrangements	remove
00101030	PatientWeight	keep
00400243	PerformedLocation	remove
00400241	PerformedStationAET	keep
00404030	PerformedStationGeoLocCodeSeq	keep
00400242	PerformedStationName	keep
00404028	PerformedStationNameCodeSeq	keep

Reproduced from https://wiki.cancerimagingarchive.net

00081052	PerformingPhysicianIdSeq	remove
00081050	PerformingPhysicianName	remove
00400250	PerformProcedureStepEndDate	incrementdate
00401102	PersonAddress	remove
00401101	PersonIdCodeSequence	remove
0040a123	PersonName	empty
00401103	PersonTelephoneNumbers	remove
40080114	PhysicianApprovingInterpretation	remove
00081048	PhysicianOfRecord	remove
00081049	PhysicianOfRecordIdSeq	remove
00081062	PhysicianReadingStudyIdSeq	remove
00402016	PlaceOrderNumberOfImagingServiceReq	empty
00181004	PlateID	keep
00400254	PPSDescription	keep
00400253	PPSID	remove
00400244	PPSStartDate	incrementdate
00400245	PPSStartTime	keep
001021c0	PregnancyStatus	keep
00400012	PreMedication	keep
Group	privategroups	keep
00131010	ProjectName	always
00181030	ProtocolName	keep
00540016	Radiopharmaceutical Information Sequence	process
00181078	Radiopharmaceutical Start DateTime	incrementdate
00181079	Radiopharmaceutical Stop DateTime	incrementdate
00402001	ReasonForImagingServiceRequest	keep
00321030	ReasonforStudy	keep
04000402	RefDigitalSignatureSeq	remove
30060024	ReferencedFrameOfReferenceUID	hashuid

00380004	ReferencedPatientAliasSeq	remove
00080092	ReferringPhysicianAddress	remove
00080090	ReferringPhysicianName	empty
00080094	ReferringPhysicianPhoneNumbers	remove
00080096	ReferringPhysiciansIDSeq	remove
00404023	RefGenPurposeSchedProcStepTransUID	hashuid
00081140	RefImageSeq	remove
00081120	RefPatientSeq	remove
00081111	RefPPSSeq	remove
00081150	RefSOPClassUID	keep
04000403	RefSOPInstanceMACSeq	remove
00081155	RefSOPInstanceUID	hashuid
00081110	RefStudySeq	remove
00102152	RegionOfResidence	remove
300600c2	RelatedFrameOfReferenceUID	hashuid
00400275	RequestAttributesSeq	remove
00321070	RequestedContrastAgent	keep
00401400	RequestedProcedureComments	keep
00321060	RequestedProcedureDescription	keep
00401001	RequestedProcedureID	remove
00401005	RequestedProcedureLocation	remove
00321032	RequestingPhysician	remove
00321033	RequestingService	remove
00102299	ResponsibleOrganization	remove
00102297	ResponsiblePerson	remove
40084000	ResultComments	keep
40080118	ResultsDistributionListSeq	remove
40080042	ResultsIDIssuer	remove
300e0008	ReviewerName	remove

00404034	ScheduledHumanPerformersSeq	remove
0038001e	ScheduledPatientInstitutionResidence	remove
0040000b	ScheduledPerformingPhysicianIDSeq	remove
00400006	ScheduledPerformingPhysicianName	remove
00400001	ScheduledStationAET	keep
00404027	ScheduledStationGeographicLocCodeSeq	keep
00400010	ScheduledStationName	keep
00404025	ScheduledStationNameCodeSeq	keep
00321020	ScheduledStudyLocation	keep
00321021	ScheduledStudyLocationAET	keep
00321000	ScheduledStudyStartDate	incrementdate
00080021	SeriesDate	incrementdate
0008103e	SeriesDescription	keep
0020000e	SeriesInstanceUID	hashuid
00080031	SeriesTime	keep
00380062	ServiceEpisodeDescription	keep
00380060	ServiceEpisodeID	remove
00131013	SiteID	SITEID
00131012	SiteName	SITENAME
001021a0	SmokingStatus	keep
00181020	SoftwareVersion	keep
00080018	SOPInstanceUID	hashuid
00082112	SourceImageSeq	remove
00380050	SpecialNeeds	keep
00400007	SPSDescription	keep
00400004	SPSEndDate	incrementdate
00400005	SPSEndTime	keep
00400011	SPSLocation	keep
00400002	SPSStartDate	incrementdate

00400003	SPSStartTime	keep
00081010	StationName	remove
00880140	StorageMediaFilesetUID	hashuid
30060008	StructureSetDate	incrementdate
00321040	StudyArrivalDate	incrementdate
00324000	StudyComments	keep
00321050	StudyCompletionDate	incrementdate
00080020	StudyDate	incrementdate
00081030	StudyDescription	keep
00200010	StudyID	empty
00320012	StudyIDIssuer	remove
0020000d	StudyInstanceUID	hashuid
00080030	StudyTime	keep
00200200	SynchronizationFrameOfReferenceUID	hashuid
0040db0d	TemplateExtensionCreatorUID	hashuid
0040db0c	TemplateExtensionOrganizationUID	hashuid
40004000	TextComments	remove
20300020	TextString	remove
00080201	TimezoneOffsetFromUTC	remove
00880910	TopicAuthor	remove
00880912	TopicKeyWords	remove
00880906	TopicSubject	remove
00880904	TopicTitle	remove
00081195	TransactionUID	hashuid
00131011	TrialName	PROJECTNAME
0040a124	UID	hashuid
Group	unspecifiedelements	keep
0040a088	VerifyingObserverIdentificationCodeSeq	remove
0040a075	VerifyingObserverName	empty

Reproduced from https://wiki.cancerimagingarchive.net

0040a073	VerifyingObserverSequence	remove
0040a027	VerifyingOrganization	remove
00384000	VisitComments	keep

More Details regarding TCIA de-identification may be found at the following links:

https://wiki.cancerimagingarchive.net/display/Public/De-identification+Knowledge+Base

https://wiki.cancerimagingarchive.net/display/Public/De-Identification+Rules