Standardized representation of the LIDC annotations using DICOM

Andrey Fedorov[*1], Matthew Hancock[2], David Clunie[3], Mathias Brockhausen[4], Jonathan Bona[4], Justin Kirby[5], John Freymann[5], Steve Pieper[6], Hugo Aerts[1,7], Ron Kikinis[1,8,9], Fred Prior[4]*

[1] *Brigham and Women's Hospital, Boston, MA*
[2] *Florida State University, Tallahassee, FL*
[3] *PixelMed Publishing, Bangor, PA*
[4] *University of Arkansas for Medical Sciences, Little Rock, AR*
[5] *Frederick National Laboratory for Cancer Research, Frederick, MD*
[6] *Isomics Inc., Cambridge, MA*
[7] *Dana Farber Cancer Institute, Boston, MA*
[8] *Fraunhofer MEVIS, Bremen, Germany*
[9] *Mathematics/Computer Science Faculty, University of Bremen, Bremen, Germany*
* *Corresponding author andrey.fedorov@gmail.com*
Google doc link: http://bit.ly/2S1DLsQ

Abstract

The Lung Imaging Data Consortium and Image Database Resource Initiative (LIDC) conducted a multi-site reader study that produced a comprehensive database of Computed Tomography (CT) scans for over 1000 subjects annotated by multiple expert readers. The result is hosted in the LIDC-IDRI collection of The Cancer Imaging Archive (TCIA). Annotations that accompany the images of the collection are stored using project-specific XML representation. This complicates their reuse, since no general-purpose tools are available to visualize or query those objects, and makes harmonization with other similar type of data non-trivial. To make the LIDC dataset more FAIR (Findable, Accessible, Interoperable, Reusable) to the research community, we prepared their standardized representation using the Digital Imaging and Communications in Medicine (DICOM) standard. This manuscript is intended to serve as a companion to the dataset to facilitate its reuse.

Background and Summary

Importance of publicly available curated databases of images for the development of novel image analysis techniques has been widely recognized for decades. The need for such collections has become particularly prominent with the recent advancement of algorithms and hardware capabilities to support computational approaches relying on deep neural networks, and the emergence of artificial intelligence (AI) methods. To be useful, such image collections need to be curated, i.e., organized and annotated in such a way that enables their use for training and evaluation of the AI systems. ImageNet is a prominent example of such database[1], which fueled the breakthrough of the deep learning revolution in the 2010s. ImageNet database has been carefully curated since 2010, and currently contains references to over 14M images. Each of those images is annotated (by humans) using WordNet hierarchy to describe the objects present in images, with over 1M of those images including bounding boxes localizing the identified objects.

It is estimated that medical imaging generates millions of clinical scans annually in the US alone[2]. Few of those scans become available for training of the AI systems, and even fewer are accompanied by annotations (labels localizing imaging findings and structured metadata describing various aspects of the disease and the imaged subject). Lack of such curated datasets has been acknowledged as a major bottleneck, if not the biggest challenge in the field of deep learning as applied to medical imaging[2,3]. It would be wrong, however, to claim that no curated medical imaging collections exist. A prominent example of such collection is the one generated by the Lung Imaging Database Consortium and Image Database Resource Initiative (LIDC/IDRI, further

referred to as LIDC), which has been a major effort supported by the National Cancer Institute (NCI) to establish a publicly available reference database of Computed Tomography (CT) images for detection, classification and quantitative assessment of lung nodules[4–6]. In an effort spanning multiple years, LIDC collaboration involved seven academic centers and eight medical imaging companies to collect a multi-site collection of CT scans for over 1000 subjects annotated by four experienced thoracic radiologists to both localize and characterize identified nodules. The resulting collection consists of the CT images stored using Digital Imaging and Communications in Medicine (DICOM) format and annotations in an XML format that follows a project-specific schema. To increase visibility and facilitate access to the resulting collection, it has been published using the resources of The Cancer Imaging Archive (TCIA)[7].

The choice of representation for storing the resulting annotations was developed for convenience during the data collection process. It proved to be effective when implemented to support radiologists annotating the data using custom software tools designed specifically for the project. Reuse of those XML annotations outside of the Consortium is more complicated. No publicly available tools were provided to accompany the dataset to either consume the annotations and support their visualization, or to provide conversion of the contours into formats supported by the platforms commonly used by imaging researchers. The representation is not self-contained, requiring the consumer of the annotations to carefully examine accompanying documentation to understand the conventions used in labeling of the nodules and the meaning of codes used for nodule characterization. This project-specific XML format makes it challenging to harmonize the annotations with the imaging data, annotations and analysis results generated for other projects within TCIA to support search and query of the data.

Despite all the challenges above, the dataset proved to be of high value and has been widely used by the community. The accompanied publications describing the dataset[4–6] accumulated over 1000 citations according to Google Scholar, and were used in a number of image analysis challenges[8]. Several tools (some of which have been released publicly) have also been contributed by the community to enable conversion of the XML annotations into alternative representations and to support exploration of the content[9–12]. Nevertheless, the XML annotations remain the only representation accessible to the users of the TCIA LIDC-IDRI collection.

The goal of the present project was to generate a standardized DICOM representation of the annotation results. There are several advantages of such representation as compared to a project-specific format. As the primary general advantage, such representation is better positioned to meet FAIR (Findable Accessible Interoperable Reusable) guiding principles for scientific data management and stewardship[13]. Beyond the benefits of standardized representation for a single dataset, this approach enables harmonization of the annotations of this specific dataset with conceptually similar results of analysis available for other collections of TCIA. As a result, aggregate queries across collections and across data types become possible, at least in principle. It also becomes easier to extend the dataset with new types of data. As an example, the same mechanisms for data encoding could be used for augmentation of the images and nodule annotations with the radiomics features derived from the nodule regions. This work utilizes tools developed earlier for interpreting XML annotations of LIDC[12] and for generating the standardized representations for image analysis results[14]. The dataset produced as a result of this work is harmonized with other standardized collections already in TCIA[15].

Methods

Introduction of the overall approach

An understanding of the content of XML annotations produced by the LIDC initiative can be gained through the peer-reviewed manuscripts published by the initiative[4–6], and the documentation available at the TCIA

LIDC-IDRI collection page[16]. Briefly, the initiative distinguished between the three groups of findings, as defined by Armato et al. [6]: *"(1) "nodules ≥ 3 mm" (defined as any lesion considered to be a nodule with greatest in-plane dimension in the range 3–30 mm regardless of presumed histology); (2) "nodules < 3 mm" (defined as any lesion considered to be a nodule with greatest in-plane dimension less than 3 mm that is not clearly benign); and (3) "non-nodules ≥ 3 mm" any other pulmonary lesion, such as an apical scar, with greatest in-plane dimension greater than or equal to 3 mm that does not possess features consistent with those of a nodule)"*. Each of the four radiologists independently reviewed all of the scans in a "blinded" phase to identify all of the findings from the three groups above. For each finding identified by a given radiologist as a *"nodule ≥ 3 mm"*, outlines were constructed in each slice where that nodule appear, while for the other two categories only the approximate center of mass was annotated. In the subsequent "unblinded" read phase each radiologist had access to the categories assigned and annotations for the nodules, and *"a radiologist's own marks then could be left unchanged, deleted, switched in terms of lesion category, or additional marks could be added"*[6]. After the unblinded phase each radiologist assessed subjective characteristics of "nodules < 3 mm", such as spiculation, subtlety, etc (discussed further).

We limited the scope of the conversion to include only "nodules ≥ 3 mm". For those nodules the annotations contained the following:

1) Planar contours defining "inclusion" or "exclusion" regions of a nodule in a given image from the CT series, organized in groups corresponding to the individual nodules. Those contours are defined as a list of coordinates defined in the space of image pixels, and corresponding to the pixels just outside the nodule (i.e., the contour pixels themselves should not be treated as belonging to the nodule).
2) Coded attributes describing various characteristics of the nodule such as opacity, conspicuity, etc.

Our general approach to standardized encoding of the data above utilized existing DICOM object definitions. A DICOM Segmentation object (SEG)[17] is the standard way to encode segmentations defined as labeled image voxels. DICOM Structured Reporting provides a versatile template TID 1500 (SR-TID1500) for communicating image-based measurements[18], both quantitative and qualitative evaluations.

Compared to a project-specific XML representation, DICOM representation offers the following advantages (also described elsewhere[19]):

- As any DICOM object, it is uniquely identified by SOPInstanceUID, and it is suitable for storage side by side with the DICOM CT dataset, and can be archived, queried and retrieved using standard DICOM storage capabilities.
- Attributes of the composite context (patient identification and attributes such as gender and age, unique identifiers for the study) are included directly in the object in the standard locations using well-defined encoding conventions.
- The use of generic, standard DICOM objects increases the possibility of using this data with general-purpose tools. A number of open source and commercial tools already include support both for SEG and SR-TID1500.
- There is a standard procedure for converting DICOM content into XML or JSON representation.
- Although limited, tools do exist for automatic validation of the DICOM objects.

DICOM SEG offers a number of desirable features for encoding segmentation results. SEG belongs to the family of DICOM enhanced multiframe objects, which means that all of the slices of the segmentation are stored in a single file. The semantics of the segmentation is encoded in a standard location, and can be described using codes from existing terminologies. References to the images being segmented can be included directly in the SEG, making it easier to trace provenance of the object and automatically retrieve the

segmented image by the visualization tools. There is a standard location to prescribe recommended color for visualization of the segmentation overlay, which is particularly helpful in situations where multiple contours for a single finding are available, as is the case in the TCIA LIDC-IDRI collection.

DICOM Structured Reporting[20] uses the key-value pairs, the "DICOM tags", to encode higher level abstraction of a tree of content, where nodes of the tree and their relationships are formalized by the DICOM Structured Reporting object definition. SR-TID1500 is just one of many templates that define constraints on the structure of the tree for the specific task of image-based measurements. DICOM SR is rooted in terminologies and codes, to deliver *structured* content. Codes are used both for defining the concepts and values assigned to those concepts. Measurements, as defined by SR-TID1500, include coded concepts corresponding to the quantity being measured, numeric value accompanied by the coded units, or coded categorical or qualitative value. In SR-TID1500, measurement is more than just the quantity and result of measurement. It is accompanied by rich context that helps interpret and reuse that measurement. Measurements derived from segmentations can reference (using unique identifiers of the respective objects) the segmentation defining the region and the image segmented. The measurement can also be accompanied by the coded data describing such attributes as finding type and location.

All of the CT images and XML annotations generated by the LIDC initiative were publicly available since 2015[16]. The data has been de-identified and curated per standard operating procedures of TCIA[7,21] to ensure no identifiable subject information is included.

The conversion process was implemented as a python script parameterizing and executing individual converters as needed. The source code of the conversion script, accompanying Jupyter Notebook, and other related items are available at https://github.com/QIICR/lidc2dicom.

Encoding of nodule annotations

Our approach to creating DICOM SEG representation of the nodule outlines was to use existing tools to enable the conversion process.

First, this work leveraged the *pylidc* python package (https://pylidc.github.io/) introduced by Hancock and Magnan[12] for accessing the volume-reconstructed annotation contours for the individual scans and subjects, as extracted from the DICOM and XML components of the TCIA LIDC-IDRI collection. *pylidc* provides an interface for iterating and querying various entities of the collection and their attributes. It also reconstructs filled multi-slice segmentations from the per-slice annotation contours. The resulting segmentations are represented as 3-dimensional *numpy* arrays, which can be padded to the dimensions of the CT image. The resulting array can be reoriented and augmented with the resolution and geometric position to construct a fully defined volume in the frame of reference of the source CT series. We utilized the Plastimatch software[22] (http://plastimatch.org/) for generating volume reconstructions of the CT scans.

```
{
  "SeriesNumber": "3000611",
  "SeriesDescription": "Segmentation of Nodule 1 - Annotation
Nodule 001",
  "ClinicalTrialTimePointID": "1",
  "ContentDescription": "Lung nodule segmentation",
  "InstanceNumber": "1",
  "ContentLabel": "SEGMENTATION",
  "ClinicalTrialCoordinatingCenterName": "TCIA",
  "segmentAttributes": [
    [
      {
        "labelID": 1,
        "SegmentDescription":
           "Nodule 1 - Annotation Nodule 001",
        "SegmentLabel": "Nodule 1 - Annotation Nodule 001",
        "recommendedDisplayRGBValue": [
          128,
          174,
          128
        ],
        "AnatomicRegionSequence": {
          "CodeMeaning": "Lung",
          "CodingSchemeDesignator": "SRT",
          "CodeValue": "T-28000"
        },
        "SegmentAlgorithmType": "MANUAL",
        "SegmentedPropertyCategoryCodeSequence": {
          "CodeMeaning": "Morphological Abnormal Structure",
          "CodingSchemeDesignator": "SRT",
          "CodeValue": "M-01000"
        },
        "SegmentedPropertyTypeCodeSequence": {
          "CodeMeaning": "Nodule",
          "CodingSchemeDesignator": "SRT",
          "CodeValue": "M-03010"
        },
        "TrackingIdentifier": "Nodule 1",
              "TrackingUniqueIdentifier":
"2.25.499735549556702272537304242512206745122406722153570861124 50"
      }
    ]
  ],
  "BodyPartExamined": "LUNG",
  "ContentCreatorName": "Reader1",
  "ClinicalTrialSeriesID": "Session1"
}
```


Recommended color for rendering the segmentation overlay

Coded anatomic region of the finding

Coded broad category of the segmentation

Coded specific type of the item being segmented

Label of the nodule and unique identifier

Figure 1: Example JSON file used to parameterize conversion of a nodule annotation into DICOM SEG representation. Coded items are defined as triplets of (CodeMeaning, CodingSchemeDesignator, CodeValue), where "SRT" denotes SNOMED-CT as the coding scheme.

Given the fully defined geometry of the segmentation in the frame of reference of the CT image, we first saved the resulting volume into NRRD format using ITK python package (https://itkpythonpackage.readthedocs.io/), and then utilized *itkimage2segimage* tool from the *dcmqi* library[14] (https://github.com/qiicr/dcmqi) to generate standard DICOM SEG representation of the annotation. The DICOM SEG conversion is parameterized by two components. First, source CT instances are used to propagate composite context information and to populate the references to the source images in the result. Second, metadata describing the segmentation is populated using a schema-constrained JSON file. An example of such file is shown in Fig.1, while all of the JSON files for each of the annotations is available alongside the converted data (see Data Citation). Semantics of the segmentation is defined by the SNOMED-CT codes assigned to segmentation category (chosen from the list of codes defined in DICOM CID 7150, and always set to "Morphologically altered structure" http://snomed.info/id/49755003) and type (DICOM CID 7151, always set to "Nodule",

http://snomed.info/id/27925004), and anatomic region (selected from the codes in DICOM CID 4, always set to "Lung", http://snomed.info/id/39607008).

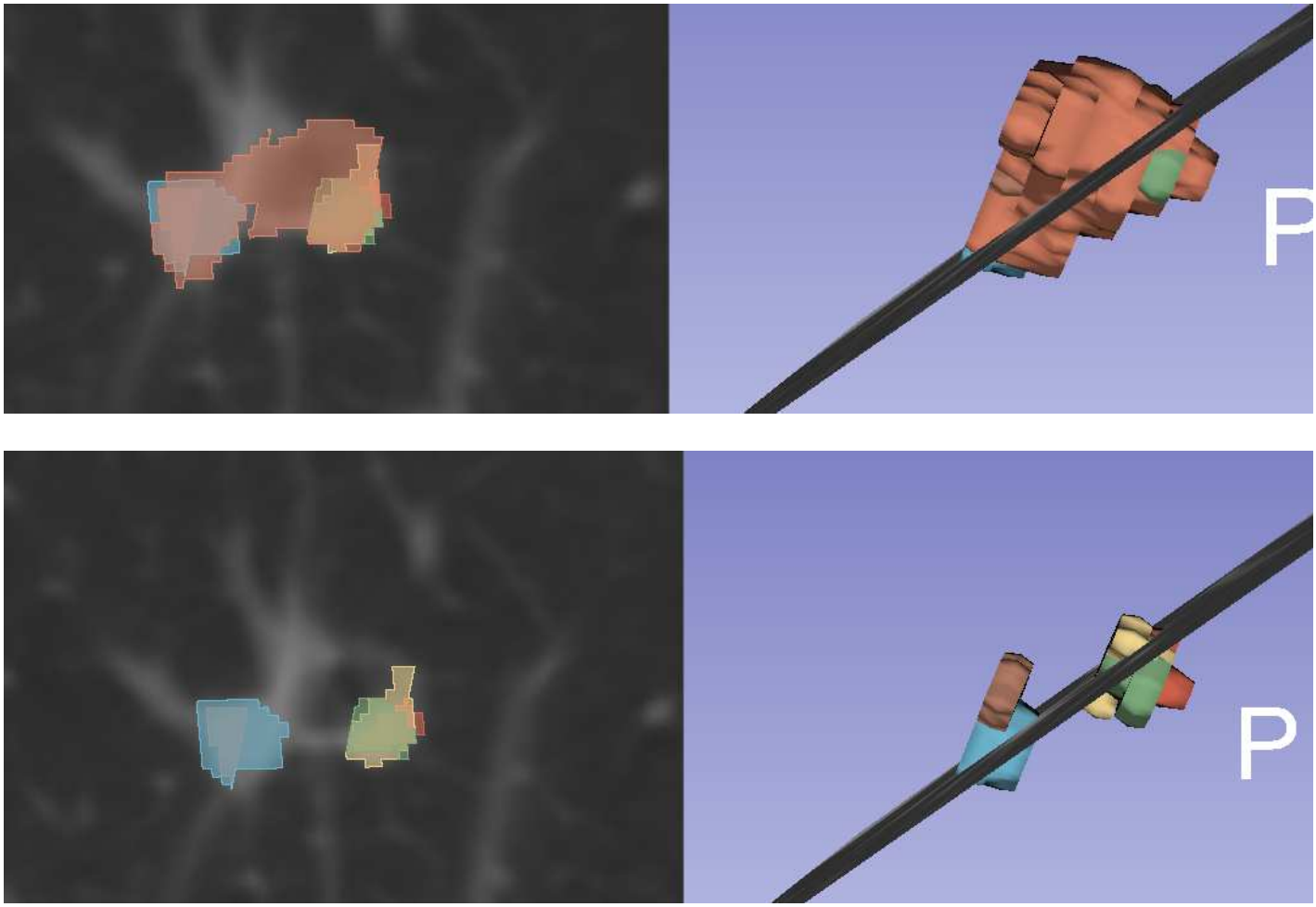

Figure 2: Illustration of annotations for subject LIDC-IDRI-0055 where in one instance (top) nodule was segmented as a single continuous structure, while other sets of annotations appear to segment separate components of the nodule (bottom). Lacking nodule or reader identifiers in the original LIDC/IDRI XML annotations, it is impossible to ascertain whether annotations shown in the bottom figure correspond to two separate nodules, or to a single nodule. All of the annotations were assigned to the same cluster by pylidc and were encoded as belonging to the same nodule in the DICOM SEG representation.

Most of the nodules were annotated by more than one expert reader. Assignment of one or more nodule-annotations to a nodule was intentionally not captured in the LIDC XML, but can be helpful in the analysis of those annotations. We utilized automatic clustering of the annotations into groups corresponding to distinctive lung nodules. The method implemented in *pylidc* clusters all nodule annotations for a given scan by computing a distance measure between the annotations. Assignment of an annotation to a given nodule is reflected in the SeriesDescription, SegmentDescription and SegmentLabel attributes. In addition, each of the annotations that were clustered to the same nodule are assigned identical and unique TrackingUniqueIdentifier values. Note that identity of the reader was intentionally not captured by the LIDC initiative. As such, it is impossible to ascertain whether any two annotations were performed by the same reader. Furthermore, some of the nodules could be interpreted as having multiple components by one reader, but might have been annotated as a single nodule by another reader (e.g., see Fig.2).

Distinctive high contrast colors were assigned to the annotations of the same nodule to facilitate simultaneous visualization of multiple annotations.

Overall summary of the decisions made and conventions followed:

- **Only nodules that were contoured volumetrically are considered.** I.e., nodules that were less than the threshold used in the LIDC study, or which had only the center identified, were not processed.
- **Each individual segmentation of a nodule is saved as a separate DICOM segmentation image instance.** It is impossible to identify all annotations done by the same reader for a given scan. If that was possible, and all nodules were annotated by the same reader during the same session, we could save all of those into the same instance. Therefore, the total number of the segmentation series per individual LIDC subject is equal to the total count of all annotations done by all readers. At the same time, visualization of the segmentations could have been more convenient were all segmentations of a nodule stored in a single DICOM object. This could be explored in the future to further augment the dataset.
- **Each segmentation instance is assigned SeriesDescription (and matching SegmentDescription and SegmentLabel attributes) to follow the convention "Nodule <nodule number> - Annotation <annotation ID>".** Nodule number is a consecutive number assigned as provided by pylidc, which uses spatial clustering of the individual annotations for a given scan to associate those to the same nodule. <annotation ID> is the identifier as assigned to the individual annotations in the XML annotation files.
- **Each nodule is assigned unique tracking identifier, which is stored in the DICOM TrackingUID attribute per segment.** This allows to link annotations corresponding to the same nodule. The same tracking identifier is used in the DICOM SR TID1500 Measurement group containing the qualitative assessments and measurements.
- **Nodule semantics is initialized uniformly for all nodules.** The following standard attributes of the DICOM Segment Description Macro were initialized as follows, to facilitate aggregation and management of segmentations across TCIA collections:
  - AnatomicRegionSequence: ("M-28000", "SRT", "Lung")
  - SegmentedPropertyCategoryCodeSequence: ("M-01000","SRT","Morphological Abnormal Structure")
  - SegmentedPropertyTypeCodeSequence: ("M-03010","SRT","Nodule")
- **Segmentation overlay display colors.** To make visualization of the nodule segmentations more user-friendly, individual annotations for a given nodule were assigned distinct, prominent colors via the RecommendedDisplayCIELabValue attribute assigned for the individual segments. The same set of colors was used for the individual annotations for individual nodules (the colors were taken from the 3D Slicer GenericColors color table). No implications about the relationship among any of the annotations that use the same colors (e.g., that they were done by the same reader) should be made: the color is used purely for facilitating visualization of the annotations overlay.
- **Empty frames were not included in the DICOM Segmentation objects.** This was a practical decision to reduce the overall disk footprint of the downloaded collection.

Encoding of annotation-derived characterizations and measurements

In the LIDC study all of the "nodules ≥ 3 mm" were subjectively assessed to describe characteristics of the nodule such as subtlety, internal structure, spiculation, lobulation, shape, sphericity, solidity, margin, and likelihood of malignancy [5,6]. For each of those characteristics, a numeric score or category was assigned, and stored in the LIDC XML representation. Explanation of the meaning of those scores or categories was provided in a separate explanatory document accompanying the XML annotations, and available on the TCIA LIDC-IDRI collection page [16]. Lack of self-contained description of the score meaning creates at least a potential for accidental reversal of the ratings.

In order to generate DICOM SR-TID1500 representation of those characterizations it was required to define codes corresponding to the concepts and values used in the process of the annotation. The original LIDC effort did not utilize the standard codes. Therefore, we first made an attempt to locate codes corresponding to the project-specific concepts and values assigned to those concepts in the existing terminologies and ontologies. Where possible, existing lexicons have been reviewed. This review included NCI Thesaurus[23], RadLex™[24] and the subset of Systematized Nomenclature of Medicine (SNOMED®)[25] codes included in the DICOM standard. To identify matching codes, we used BioPortal[26] (http://bioportal.bioontology.org/), the Ontology Lookup Service[27] (https://www.ebi.ac.uk/ols), and the RadLex Term Browser (http://www.radlex.org/). We have also utilized the terms defined by the Imaging Biomarker Standardization Initiative (IBSI)[28], which defines those in the context of radiomics feature extraction. Furthermore, we consulted the earlier report by Opulencia et al.[29] mapping LIDC concepts to RadLex and refined our selection accordingly. Where matches were identified, standard codes were used. However, if the match was deemed to have the potential of losing the project-specific meaning, we opted for introducing non-standard codes. The non-standard new codes can be identified by the "99LIDCQIICR" coding scheme (prefix "99" is the DICOM-defined means of flagging a coding scheme as non-standard). The codes for the resulting concepts and values are summarized in Tables 1 and 2, respectively. Highlighted entries correspond to the terms that were identified after the release of the first version of the dataset (and the first version of the preprint).

Table 1. LIDC-IDRI Evaluation Concepts

Coding Scheme Designator	Code Value	Code Meaning	LIDC-IDRI concept name verbatim	LIDC definition
NCIt	C45992	Subtlety Score	Subtlety	Radiologist assessment of nodule subtlety on 1-5 scale
99LIDCQIICR	200	Internal structure	Internal structure	Radiologist assessment of nodule internal structure
NCIt	C3672	Calcification	Calcification	Radiologist assessment of internal calcification of nodule
IBSI	QCFX	Sphericity	Sphericity	Radiologist assessment of shape of nodule in terms of its roundness/sphericity with only 3 terms defined
NCIt	C25563	Margin	Margin	Radiologist assessment of

				nodule margin on a 1-5 scale with only the extreme values explicitly defined
NCIt	C62175	Lobular Pattern	Lobulation	Radiologist assessment of nodule lobulation on a 1-5 scale with only the extreme values explicitly defined
NCIt	C28749	Spiculate (synonym: Spiculation)	Spiculation	Radiologist assessment of nodule spiculation on a 1-5 scale with only the extreme values explicitly defined
NCIt	C41144	Texture	Texture	Radiologist assessment of nodule internal texture with only 3 terms defined
RadLex	RID36042	Malignant neoplasm (synonym: Malignancy)	Likelihood of malignancy	Radiologist subjective assessment of likelihood of malignancy of this nodule (ASSUMING 60-year-old male smoker)

Table 2. LIDC-IDRI Evaluation Concept Values.

Coding Scheme Designator	Code Value	Code Meaning	LIDC-IDRI concept name verbatim
Subtlety[1]			
99LIDCQIICR	101	1 out of 5 (Extremely subtle)	1 - Extremely subtle
99LIDCQIICR	102	2 out of 5 (Moderately subtle)	2 - Moderately subtle

[1] Annotated XML file contains definitions for the subtlety score only for scores 1 and 5, while Fig.7 showing software interface in [5] contains definitions for all of the scores, as listed in the table.

99LIDCQIICR	103	3 out of 5 (Fairly subtle)	3 - Fairly subtle
99LIDCQIICR	104	4 out of 5 (Moderately obvious)	4 - Moderately obvious
99LIDCQIICR	105	5 out of 5 (Obvious)	5 - Obvious
Internal structure			
NCIt	C12471	Soft tissue	Soft tissue
NCIt	C25278	Fluid	Fluid
NCIt	C12472	Adipose tissue	Fat
NCIt	C73434	Air	Air
Calcification			
RadLex	RID35453	Popcorn calcification sign	1 - Popcorn appearance
99LIDCQIICR	302	Laminated appearance	2 - Laminated appearance
RadLex	RID5741	Solid	3 - Solid appearance
99LIDCQIICR	304	Non-central appearance	4 - Non-central appearance
RadLex	RID5827	Central	5 - Central calcification
RadLex	RID28473	Absent	6 - Absent
Sphericity			
RadLex	RID5811	linear	1 - Linear appearance
99LIDCQIICR	002	2 out of 5	2
RadLex	RID5800	ovoid	3 - Ovoid appearance
99LIDCQIICR	004	4 out of 5	4
RadLex	RID5799	round	5 - Round appearance
Margin			
RadLex	RID5709	Indistinct margin (synonym: poorly defined margin)	1 - Poorly defined
99LIDCQIICR	002	2 out of 5	2
99LIDCQIICR	003	3 out of 5	3
99LIDCQIICR	004	4 out of 5	4
RadLex	RID5707	Circumscribed margin (synonym:	5 - Sharp margin

		sharpy-defined margin)	
Lobulation			
99LIDCQIICR	601	1 out of 5 (No lobulation)	1 - No lobulation
99LIDCQIICR	002	2 out of 5	2
99LIDCQIICR	003	3 out of 5	3
99LIDCQIICR	004	4 out of 5	4
99LIDCQIICR	605	5 out of 5 (Marked lobulation)	5 - Marked lobulation
Spiculation			
99LIDCQIICR	701	1 out of 5 (No spiculation)	1 - No spiculation
99LIDCQIICR	002	2 out of 5	2
99LIDCQIICR	003	3 out of 5	3
99LIDCQIICR	004	4 out of 5	4
99LIDCQIICR	705	5 out of 5 (Marked spiculation)	5 - Marked spiculation
Texture			
RadLex	RID50153	Non-solid pulmonary nodule (synonym: pure ground-glass pulmonary nodule)	1 - Non-solid/Ground Glass Opacity
99LIDCQIICR	002	2 out of 5	2
RadLex	RID50152	part-solid pulmonary nodule	3 - Part-solid/mixed
99LIDCQIICR	004	4 out of 5	4
RadLex	RID50151	solid pulmonary nodule	5 - Solid texture
Malignancy[2]			
99LIDCQIICR	901	1 out of 5 (Highly Unlikely for Cancer)	1 - Highly Unlikely for Cancer
99LIDCQIICR	902	2 out of 5 (Moderately Unlikely for Cancer)	2 - Moderately Unlikely for Cancer
99LIDCQIICR	903	3 out of 5 (Indeterminate Likelihood)	3 - Indeterminate Likelihood
99LIDCQIICR	904	4 out of 5 (Moderately Suspicious for Cancer)	4 - Moderately Suspicious for Cancer

[2] Note minor inconsistencies in the definitions in the annotated XML file as compared to Fig.7 in McNitt-Gray et al.[5].

99LIDCQIICR	905	5 out of 5 (Highly Suspicious for Cancer)	5 - Highly Suspicious for Cancer

An example demonstrating the differences between the original approach used in LIDC/IDRI XML, and the use of codes for concept and values can be observed in Fig.3.

```
<characteristics>
  <subtlety>5</subtlety>
  <internalStructure>1</internalStructure>
  <calcification>6</calcification>
  <sphericity>2</sphericity>
  <margin>3</margin>
  <lobulation>3</lobulation>
  <spiculation>3</spiculation>
  <texture>4</texture>
  <malignancy>3</malignancy>
</characteristics>
```

```
C3672,NCIt,"Calcification")=(RID28473,RadLex,"Absent")>
(200,99LIDCQIICR,"Internal structure")=(C12471,NCIt,"Soft
tissue")>
(400,99LIDCQIICR,"Sphericity")=(002,99LIDCQIICR,"2 out of 5")>
(C45992,NCIt,"Subtlety score")=(105,99LIDCQIICR,"5 out of 5
(Obvious)")>
(700,99LIDCQIICR,"Spiculation")=(003,99LIDCQIICR,"3 out of 5")>
(C62175,NCIt,"Lobular Pattern")=(003,99LIDCQIICR,"3 out of 5")>
(C25563,NCIt,"Margin")=(003,99LIDCQIICR,"3 out of 5")>
(C41144,NCIt,"Texture")=(004,99LIDCQIICR,"4 out of 5")>
(RID36042,RadLex,"Malignancy")=(903,99LIDCQIICR,"3 out of 5
(Indeterminate Likelihood)")>
```

Figure 3: Comparison of the communication of the nodule characteristics using the original LIDC XML approach (left) and the code-based approach used in DICOM SR-TID1500 (right). The latter approach uses existing terminologies, where possible, and includes definitions of the values assigned to coded concepts to make the characterizations document self-contained.

In addition to the subjective characterizations, we included the measurements calculated by *pylidc* coded as follows:

- Diameter: (“M-02550”, “SRT”, “Diameter”), units: (“mm”, “UCUM”, “millimeter”)
- Surface area: (“C0JK”, “IBSI”, “Surface area of mesh”), units: (“mm2”, “UCUM”, “square millimeter”)
- Volume: (“G-D705”, “SRT”, “Volume”), units: (“mm3”, “UCUM”, “cubic millimeter”)

Qualitative characterizations were extracted from XML and associated with the nodule annotations by *pylidc.* Generation of DICOM SR-TID1500 content was done using the *tid1500writer* tool from *dcmqi.* Similar to the process of generating DICOM SEG, the conversion process was parameterized using schema-constrained JSON that described the characterizations and measurements to be encoded (as shown in Fig.3), and associated them with the segmentations and source CT images that were used to derive them. The JSON files that were used in the process of conversion accompany the conversion results (see Data Citation).

Technical validation

Conformance of the converted objects to the DICOM standard was established using the *dciodvfy* tool from the *dicom3tools* software (http://www.dclunie.com/dicom3tools.html). Consistency of the metadata stored in the DICOM objects and the *pylidc* database was confirmed. Consistency of visualization of the segmentations between *pylidc* representation and the DICOM representation was also confirmed. Furthermore, we developed a demonstration that enables interrogation and exploration of the various types of metadata combined across the different object types via unified queries. To do that, selected metadata of the CT, SEG and SR objects was first extracted into a set of tab-delimited files using the *dcm2tables* open source tool (https://github.com/QIICR/dcm2tables).

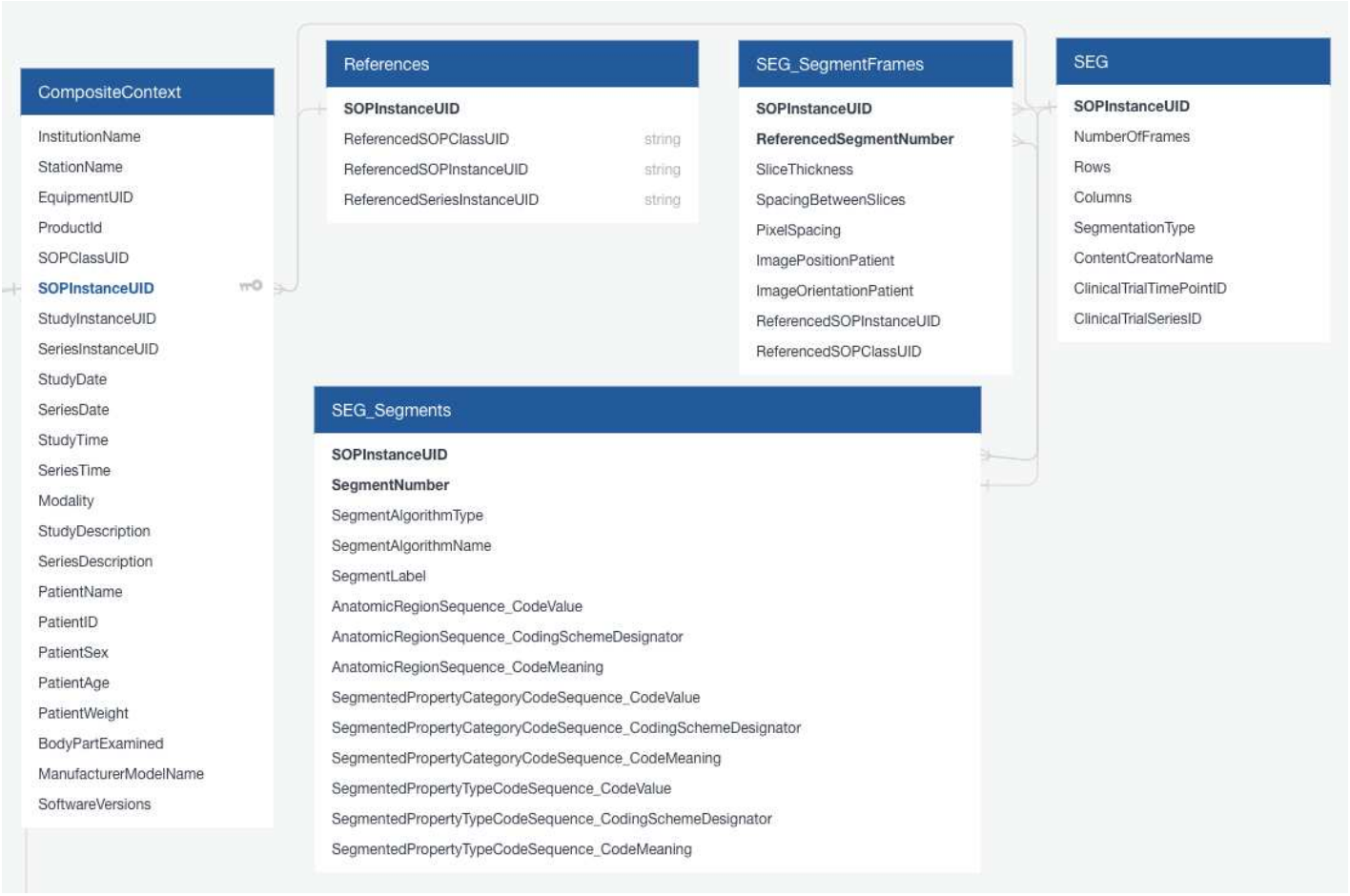

Figure 4: DICOM attributes and their assignment to the individual tables, as extracted by the dcm2tables tool, for the DICOM SEG objects.

dcm2tables extracts DICOM metadata following the tool-specific table schema which aims to mirror the hierarchical organization of DICOM metadata. As an example, “CompositeContext” table contains metadata attributes expected to be present in every DICOM object (e.g., SOPInstanceUID, SeriesDescription, StudyDate), and is linked by the SOPInstanceUID key to the “CT” table containing attributes specific to the DICOM Computed Tomography object (i.e., metadata attributes extracted from a CT object will be split between the “CompositeContext” and “CT” tables). Handling of the objects that follow hierarchical organization is more involved. Segmentation object contains references to the images being segmented, includes segment-specific metadata, and contains metadata associated with the individual frames of a given segment. Figure 4 shows the various tables that contain segmentation-specific metadata. Tables containing the metadata corresponding DICOM SR documents containing qualitative assessments and measurements for the individual annotations also follow the hierarchical organization (see Fig.5). “SR” table contains selected attributes that are expected in any DICOM SR object. DICOM SR TID1500 will contain one or more measurement groups. Some of the metadata that may be specified at that level is extracted into the “SR1500_MeasurementGroups” table. In turn, a measurement group will contain one or more measurements or qualitative evaluations, metadata associated with which is extracted into “SR1500_QualitativeEvaluations” and “SR1500_Measurements” tables, respectively.

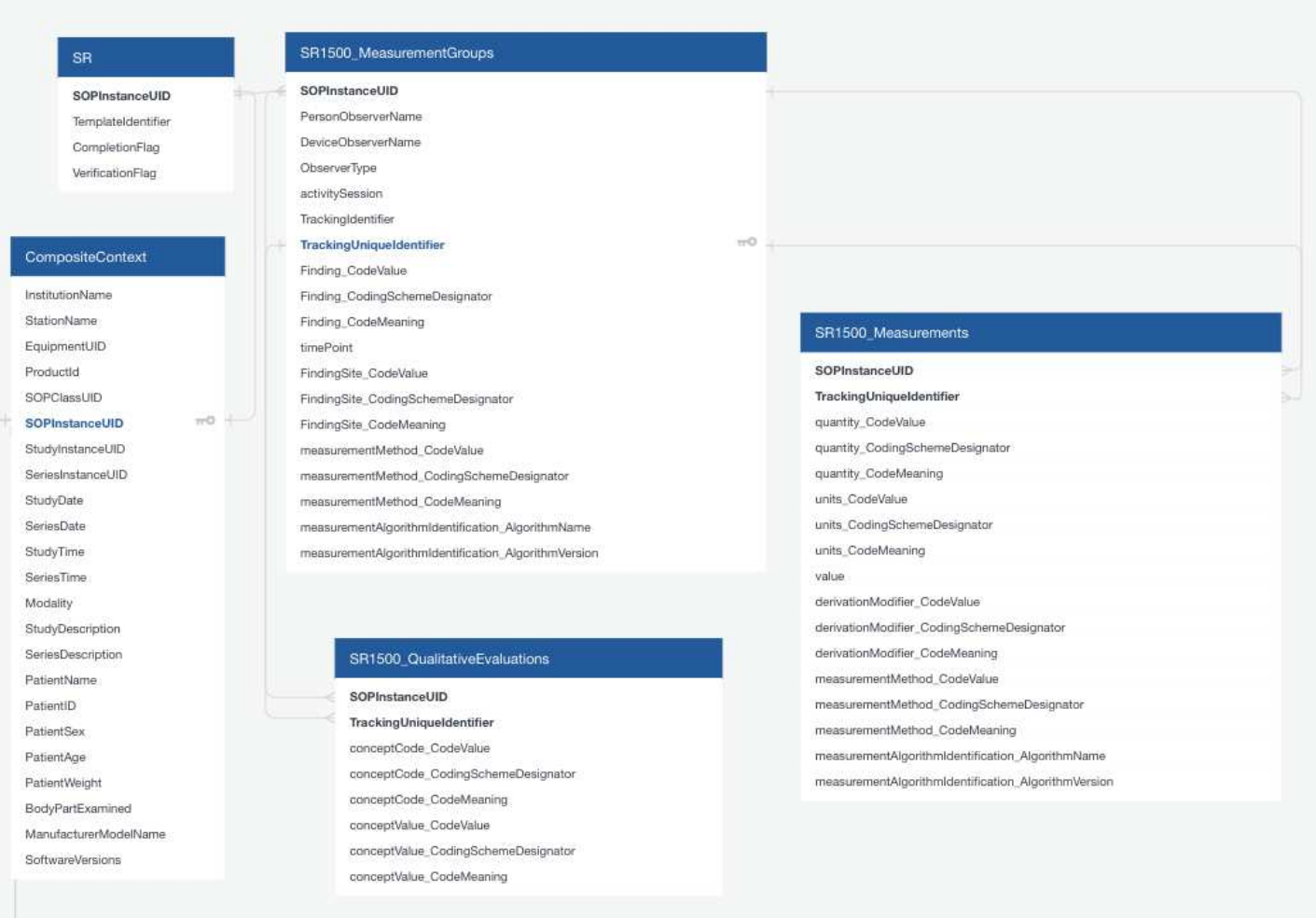

Figure 5: DICOM attributes and their assignment to the individual tables, as extracted by the dcm2tables tool, for the DICOM SR objects containing LIDC-IDRI collection annotations and measurements.

The resulting tables can be used to interrogate metadata using any relational database tools. For the purposes of our evaluation and demonstration of data exploration, we utilized the *pandas* package of *Python*, in combination with *Jupyter Notebook* and the image visualization capabilities provided by *pylidc* and *3D Slicer*[30] (https://sllicer.org).

First, consistency of the metadata between the DICOM representation and the content of *pylidc* was verified. The pseudocode of the approach used for this verification is shown in Fig. 6.

```
FOR subject in LIDC-IDRI
  FOR scan in pylidc scans for subject
    FOR nodule in pylidc nodules for scan
      FOR annotation in pylidc annotations for nodule
        Construct segmentLabel
        Locate DICOM Segment based on segmentLabel
        Confirm there is one and only one DICOM Segment with segmentLabel
        Locate measurementGroup corresponding to the DICOM Segment
        Confirm all measurements in measurementGroup are identical to those in pylidc
       Confirm all qualitative assessments in measurementGroup are identical to those in
pylidc
```

Figure 6: Pseudocode of the procedure used to confirm consistency of the qualitative assessments and

measurements assigned to the individual annotations between the pylidc and DICOM representation.

Individual annotations, or all annotations for a given nodule, can be visualized either using *pylidc* (based on *pylidc*-specific representations of data) or *3D Slicer* (using the standardized DICOM objects) (e.g., see Fig.7).

Figure 7: Visualization of the same nodule annotation in 3D Slicer (left, green overlay) and pylidc viewer (right, red outline overlay). The annotation shown corresponds to the largest nodule in the collection (LIDC-IDRI-0834 nodule 1).

Standardized metadata extracted into tabular form can be used to collect summary statistics or generate various visualizations (e.g., see Fig.8).

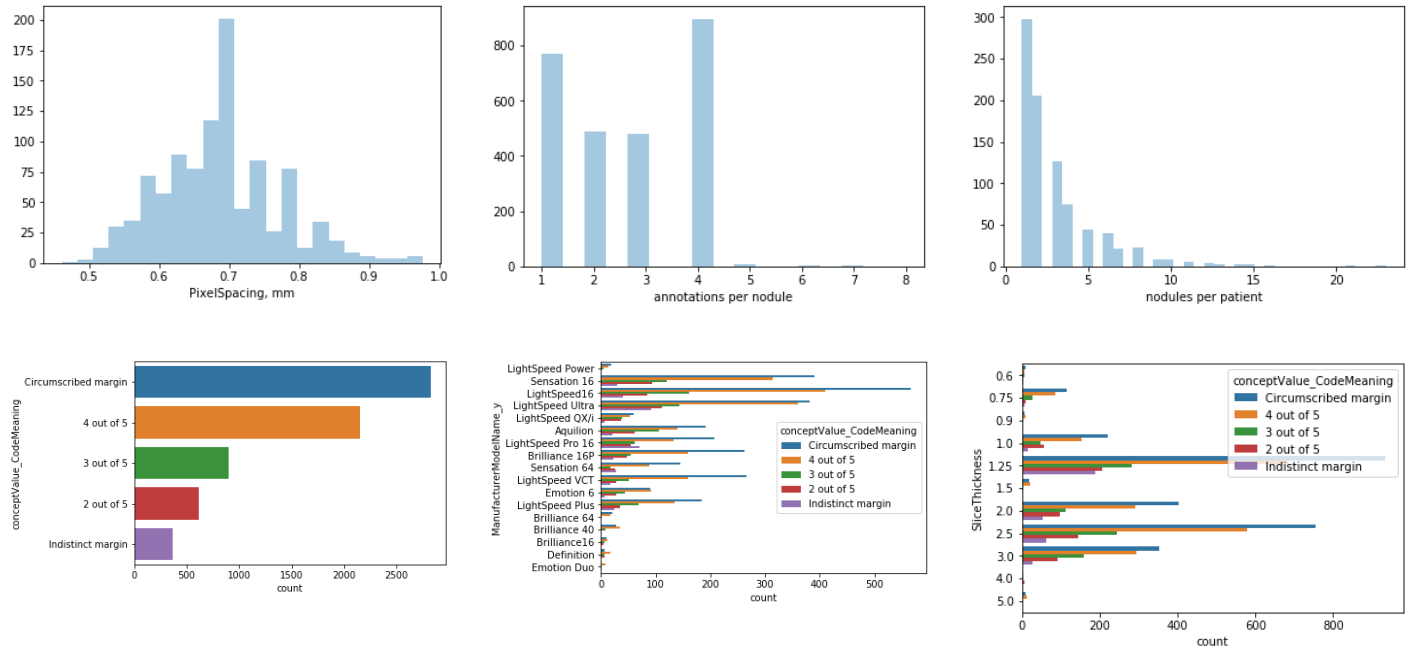

Figure 8: Example summary visualizations of various aspects of the data. First row: distribution of pixel spacing, annotations per nodule and nodules per subject. Bottom row: distribution of margin assessment categories - overall, per individual scanner manufacturers, and per individual slice thickness values.

Usage Notes

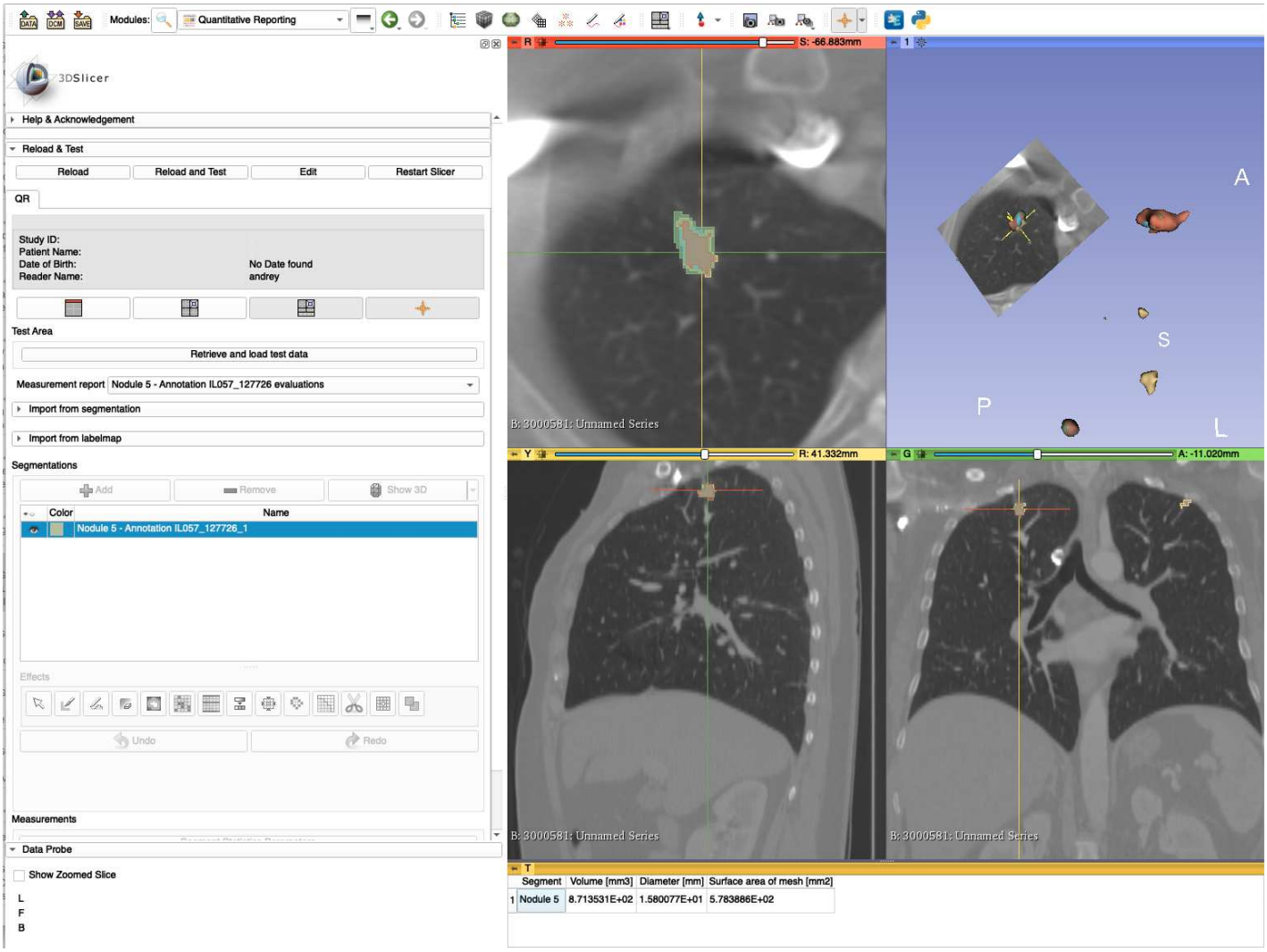

Figure 9: Example of visualization of the annotations and the associated measurements using 3D Slicer QuantitativeReporting extension. Shown is one of the CT scans and the corresponding annotations for subject LIDC-IDRI-0055.

Visualization Visualization of the DICOM SEG objects and the associated measurements can be performed using the *3D Slicer* software[30] (https://slicer.org). *QuantitativeReporting* extension[14] (https://github.com/QIICR/QuantitativeReporting) should be installed first, since support of DICOM SEG and SR-TID1500 is not available in the core application. The extension can be installed by first downloading the latest version of the 3D Slicer application package from https://download.slicer.org, and then using the *Extension Manager* to install the extension. Detailed installation instructions are available in the extension documentation. Once installed, DICOM images should be imported into the application using *DICOM Browser* module, upon which any SR object from the collection can be loaded, triggering automatic load of the corresponding SEG and CT image series.

Extraction of metadata and conversion of the DICOM objects into alternative representations can be done using a variety of tools and approaches. *dcmqi* can be used to extract the metadata specific to SEG and SR-TID1500 and store it in *dcmqi*-specific JSON representation. *segimage2itkimage* tool of *dcmqi* can be used to convert the pixel data for individual segments into commonly used volumetric formats readable by ITK

(including ITK python package, which can read those volumetric formats as *numpy* arrays) and commonly used by researchers, such as NIfTI or NRRD.

Exploration of metadata Interactive exploration of the metadata combined with the visualization of the images and annotation can be done, as an example, using Jupyter Notebooks and python *pandas* package. A notebook accompanying this manuscript is available in https://github.com/qiicr/lidc2dicom/notebooks. The examples shown in Fig. 8 were generated using this notebook.

Open source OFFIS DICOM Toolkit (DCMTK)[31] (https://dcmtk.org) provides a number of command line tools to support exploration of DICOM data and DICOM SR objects specifically.

Imaging Measurements (126010, DCM)

Measurement Group (125007, DCM)

Observation Context: Activity Session (C67447, NCIt) = "1"
Observation Context: Tracking Identifier (112039, DCM) = "Nodule 1"
Observation Context: Tracking Unique Identifier (112040, DCM) = 2.25.57872248199145978602917759351836292747440337293994762318787

Finding (121071, DCM):
Nodule (M-03010, SRT)

Observation Context: Time Point (C2348792, UMLS) = "1"

Referenced Segment (121191, DCM):
SG image

Source series for segmentation (121232, DCM):
1.3.6.1.4.1.14519.5.2.1.6279.6001.330425234131526435132846006585

Concept Modifier: Finding Site (G-C0E3, SRT) = Lung (T-28000, SRT)

Volume (G-D705, SRT):
7.963384E+03 mm3

Concept Modifier: Algorithm Name (111001, DCM) = "pylidc"
Concept Modifier: Algorithm Version (111003, DCM) = "0.2.0"

Diameter (M-02550, SRT):
3.561117E+01 mm

Concept Modifier: Algorithm Name (111001, DCM) = "pylidc"
Concept Modifier: Algorithm Version (111003, DCM) = "0.2.0"

Surface area of mesh (C0JK, IBSI):
3.547097E+03 mm2

Concept Modifier: Algorithm Name (111001, DCM) = "pylidc"
Concept Modifier: Algorithm Version (111003, DCM) = "0.2.0"

Calcification (C3672, NCIt):
Absent (RID28473, RadLex)

Internal structure (200, 99LIDCQIICR):
Soft tissue (C12471, NCIt)

Sphericity (400, 99LIDCQIICR):
round (RID5799, RadLex)

Subtlety score (C45992, NCIt):
5 out of 5 (Obvious) (105, 99LIDCQIICR)

Spiculation (700, 99LIDCQIICR):
5 out of 5 (Marked spiculation) (705, 99LIDCQIICR)

Lobulation (600, 99LIDCQIICR):
1 out of 5 (No lobulation) (601, 99LIDCQIICR)

Margin (C25563, NCIt):
4 out of 5 (004, 99LIDCQIICR)

Texture (C41144, NCIt):
4 out of 5 (004, 99LIDCQIICR)

Malignancy (900, 99LIDCQIICR):
4 out of 5 (Moderately Suspicious for Cancer) (904, 99LIDCQIICR)

Figure 10: Section of the HTML rendering of a SR-TID1500 object generated with dcr2html command line tool of DCMTK (argument +Cn was used to render codes for concept names).

```
<contains CONTAINER:(126010,DCM,"Imaging Measurements")=SEPARATE>
  <contains CONTAINER:(125007,DCM,"Measurement Group")=SEPARATE>
    <has obs context TEXT:(C67447,NCIt,"Activity Session")="1">
    <has obs context TEXT:(112039,DCM,"Tracking Identifier")="Nodule 1">
    <has obs context UIDREF:(112040,DCM,"Tracking Unique Identifier")="2.25.57872248199145978602917759351836292747440337293994762318787">
    <contains CODE:(121071,DCM,"Finding")=(M-03010,SRT,"Nodule")>
    <has obs context TEXT:(C2348792,UMLS,"Time Point")="1">
    <contains IMAGE:(121191,DCM,"Referenced Segment")=(SG image,,1)>
    <contains UIDREF:(121232,DCM,"Source series for segmentation")="1.3.6.1.4.1.14519.5.2.1.6279.6001.330425234131526435132846006585">
    <has concept mod CODE:(G-C0E3,SRT,"Finding Site")=(T-28000,SRT,"Lung")>
    <contains NUM:(G-D705,SRT,"Volume")="7.963384E+03" (mm3,UCUM,"cubic millimeter")>
      <has concept mod TEXT:(111001,DCM,"Algorithm Name")="pylidc">
      <has concept mod TEXT:(111003,DCM,"Algorithm Version")="0.2.0">
    <contains NUM:(M-02550,SRT,"Diameter")="3.561117E+01" (mm,UCUM,"millimeter")>
      <has concept mod TEXT:(111001,DCM,"Algorithm Name")="pylidc">
      <has concept mod TEXT:(111003,DCM,"Algorithm Version")="0.2.0">
    <contains NUM:(C0JK,IBSI,"Surface area of mesh")="3.547097E+03" (mm2,UCUM,"square millimeter")>
      <has concept mod TEXT:(111001,DCM,"Algorithm Name")="pylidc">
      <has concept mod TEXT:(111003,DCM,"Algorithm Version")="0.2.0">
    <contains CODE:(C3672,NCIt,"Calcification")=(RID28473,RadLex,"Absent")>
    <contains CODE:(200,99LIDCQIICR,"Internal structure")=(C12471,NCIt,"Soft tissue")>
    <contains CODE:(400,99LIDCQIICR,"Sphericity")=(RID5799,RadLex,"round")>
    <contains CODE:(C45992,NCIt,"Subtlety score")=(105,99LIDCQIICR,"5 out of 5 (Obvious)")>
    <contains CODE:(700,99LIDCQIICR,"Spiculation")=(705,99LIDCQIICR,"5 out of 5 (Marked spiculation)")>
    <contains CODE:(600,99LIDCQIICR,"Lobulation")=(601,99LIDCQIICR,"1 out of 5 (No lobulation)")>
    <contains CODE:(C25563,NCIt,"Margin")=(004,99LIDCQIICR,"4 out of 5")>
    <contains CODE:(C41144,NCIt,"Texture")=(004,99LIDCQIICR,"4 out of 5")>
    <contains CODE:(900,99LIDCQIICR,"Malignancy")=(904,99LIDCQIICR,"4 out of 5 (Moderately Suspicious for Cancer)")>
```

Figure 11: Output of the dsrdump tool for the same content as shown in Fig.10.

dcmdump can be used to examine the content of any DICOM object at the level of the individual DICOM attributes. *dsr2html* tool can be used to generate human-readable rendering of the SR-TID1500 content (an example of such rendering for one of the annotations is shown in Fig.10). Similarly, abbreviated content of the DICOM SR tree can be displayed using *dsrdump* tool (see Fig.11).

Programmatic access

```
    <contains CODE:(C3672,NCIt,"Calcification")=(RID28473,RadLex,"Absent")>
```

```
(fffe,e000) na (Item with undefined length #=4)          # u/l, 1 Item
  (0040,a010) CS [CONTAINS]                                #   8, 1 RelationshipType
  (0040,a040) CS [CODE]                                    #   4, 1 ValueType
  (0040,a043) SQ (Sequence with undefined length #=1)      # u/l, 1 ConceptNameCodeSequence
    (fffe,e000) na (Item with undefined length #=3)          # u/l, 1 Item
      (0008,0100) SH [C3672]                                   #   6, 1 CodeValue
      (0008,0102) SH [NCIt]                                    #   4, 1 CodingSchemeDesignator
      (0008,0104) LO [Calcification]                           #  14, 1 CodeMeaning
    (fffe,e00d) na (ItemDelimitationItem)                    #   0, 0 ItemDelimitationItem
  (fffe,e0dd) na (SequenceDelimitationItem)                #   0, 0 SequenceDelimitationItem
  (0040,a168) SQ (Sequence with undefined length #=1)      # u/l, 1 ConceptCodeSequence
    (fffe,e000) na (Item with undefined length #=3)          # u/l, 1 Item
      (0008,0100) SH [RID28473]                                #   8, 1 CodeValue
      (0008,0102) SH [RadLex]                                  #   6, 1 CodingSchemeDesignator
      (0008,0104) LO [Absent]                                  #   6, 1 CodeMeaning
    (fffe,e00d) na (ItemDelimitationItem)                    #   0, 0 ItemDelimitationItem
  (fffe,e0dd) na (SequenceDelimitationItem)                #   0, 0 SequenceDelimitationItem
(fffe,e00d) na (ItemDelimitationItem)                    #   0, 0 ItemDelimitationItem
```

```
  <code>
    <relationship>CONTAINS</relationship>
    <concept>
      <value>C3672</value>
```

```
      <scheme>
        <designator>NCIt</designator>
      </scheme>
      <meaning>Calcification</meaning>
    </concept>
    <value>RID28473</value>
    <scheme>
      <designator>RadLex</designator>
    </scheme>
    <meaning>Absent</meaning>
  </code>
```

Figure 12: Top: content item from the SR tree level view produced by dsrdump; middle: the same content shown at the level of individual DICOM attributes; bottom: the same content as generated by the DCMTK dsr2xml tool.

A number of toolkits are available to support interpretation of DICOM objects and access to their content at the level of individual attributes. Open source toolkits providing this functionality include DCMTK and Grassroots DICOM (GDCM) (http://gdcm.sourceforge.net/) in C++, *pydicom* in Python (https://github.com/pydicom/pydicom), dcmjs (https://github.com/dcmjs-org/dcmjs) in JavaScript, and PixelMed Java DICOM toolkit in Java (https://www.pixelmed.com/dicomtoolkit.html). Programmatic access to the metadata attributes of DICOM SEG objects should be rather straightforward with the basic understanding of the DICOM concepts. Interpretation of the DICOM SR-TID1500 is somewhat more complicated if done at the level of individual DICOM attributes. To illustrate the encoding of the DICOM SR content tree, Fig. 12 shows the section of an SR-TID1500 document for a single node of the content tree. DCMTK *dcmsr* module provides Application Programming Interface (API) that allows to iterate DICOM SR tree content. However, that API is only available in C++. *pydicom* does not provide the abstraction to iterate over the content of the SR tree.

Lacking versatile support of DICOM SR tree interrogation, a practical approach to extracting structured content into alternative representations could be to instead to interpret a tool-specific representation of the content. *dsr2xml* command line tool of DCMTK can be used to convert the content of the DICOM SR document into non-standard *dcmsr*-specific XML representation. *tid1500reader* tool from *dcmqi* will store the SR-TID1500-specific content into *dcmqi*-specific JSON representation. Although those intermediate representations are not standard, they can be generated using publicly available tools from standard DICOM representation, and can simplify programmatic interpretation of those objects lacking a more convenient API functionality in languages other than C++.

Acknowledgments

This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health under Contract No. HHSN261200800001E, and by the NIH grants U24 CA180918, U24 CA199460 and U01 CA190234. 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. Under this contract the University of Arkansas for Medical Sciences is funded by Leidos Biomedical Research subcontract 16X011.

References

1. Deng, J., Dong, W., Socher, R., Li, L., Li, K. & Fei-Fei, L. ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (ieeexplore.ieee.org, 2009). doi:10.1109/CVPR.2009.5206848
2. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18,** 500–510 (2018).

3. Syeda-Mahmood, T. Role of Big Data and Machine Learning in Diagnostic Decision Support in Radiology. *J. Am. Coll. Radiol.* **15,** 569–576 (2018).
4. Armato, S. G., III, McLennan, G., McNitt-Gray, M. F., Meyer, C. R., Yankelevitz, D., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Reeves, A. P., Croft, B. Y. & Clarke, L. P. Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community1. *Radiology* **232,** 739–748 (2004).
5. McNitt-Gray, M. F., Armato, S. G., III, Meyer, C. R., Reeves, A. P., McLennan, G., Pais, R. C., Freymann, J., Brown, M. S., Engelmann, R. M., Bland, P. H., Laderach, G. E., Piker, C., Guo, J., Towfic, Z., Qing, D. P.-Y., Yankelevitz, D. F., Aberle, D. R., van Beek, E. J. R., MacMahon, H., Kazerooni, E. A., Croft, B. Y. & Clarke, L. P. The Lung Image Database Consortium (LIDC) Data Collection Process for Nodule Detection and Annotation. *Acad. Radiol.* **14,** 1464–1474 (2007).
6. Armato, S. G., 3rd, McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Van Beeke, E. J. R., Yankelevitz, D., Biancardi, A. M., Bland, P. H., Brown, M. S., Engelmann, R. M., Laderach, G. E., Max, D., Pais, R. C., Qing, D. P. Y., Roberts, R. Y., Smith, A. R., Starkey, A., Batrah, P., Caligiuri, P., Farooqi, A., Gladish, G. W., Jude, C. M., Munden, R. F., Petkovska, I., Quint, L. E., Schwartz, L. H., Sundaram, B., Dodd, L. E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A. V., Gupte, S., Sallamm, M., Heath, M. D., Kuhn, M. H., Dharaiya, E., Burns, R., Fryd, D. S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S. & Croft, B. Y. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* **38,** 915–931 (2011).
7. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. & Prior, F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26,** 1045–1057 (2013).
8. Kalpathy-Cramer, J., Freymann, J. B., Kirby, J. S., Kinahan, P. E. & Prior, F. W. Quantitative Imaging Network: Data Sharing and Competitive AlgorithmValidation Leveraging The Cancer Imaging Archive. *Transl. Oncol.* **7,** 147–152 (2014).
9. Lin, H., Chen, Z. & Wang, W. A pulmonary nodule view system for the Lung Image Database Consortium (LIDC). *Acad. Radiol.* **18,** 1181–1185 (2011).
10. Zeng, C., Lin, H. & Wang, W. Development of a Data Integration and Visualization Software for LIDC. *JSW* (2013). at <http://www.jsoftware.us/vol8/jsw0809-28.pdf>
11. Lampert, T. A., Stumpf, A. & Gancarski, P. An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. *IEEE Trans. Image Process.* **25,** 2557–2572 (2016).
12. Hancock, M. C. & Magnan, J. F. Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods. *J Med Imaging (Bellingham)* **3,** 044504 (2016).
13. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016).
14. Herz, C., Fillion-Robin, J.-C., Onken, M., Riesmeier, J., Lasso, A., Pinter, C., Fichtinger, G., Pieper, S., Clunie, D., Kikinis, R. & Fedorov, A. dcmqi: An Open Source Library for Standardized Communication of Quantitative Image Analysis Results Using DICOM. *Cancer Res.* **77,** e87–e90 (2017).
15. Beichel, R. R., Ulrich, E. J., Bauer, C., Wahle, A., Brown, B., Chang, T., Plichta, K. A., Smith, B. J., Sunderland, J. J., Braun, T., Fedorov, A., Clunie, D., Onken, M., Riesmeier, J., Pieper, S., Kikinis, R., Graham, M. M., Casavant, T. L., Sonka, M. & Buatti, J. M. QIN-HEADNECK - The Cancer Imaging Archive (TCIA). (2016). doi:10.7937/K9/TCIA.2015.K0F5CGLI
16. Armato, S. G., III, McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P. & Clarke, L. P. Data From LIDC-IDRI. (2015). doi:10.7937/K9/TCIA.2015.LO9QL9SX

17. National Electrical Manufacturers Association (NEMA). in *DICOM PS3.3 - Information Object Definitions* (2016). at <http://dicom.nema.org/medical/dicom/current/output/html/part03.html#sect_A.51>
18. National Electrical Manufacturers Association (NEMA). in *DICOM PS3.16 - Content Mapping Resource* (2016). at <http://dicom.nema.org/medical/dicom/current/output/chtml/part16/chapter_A.html#sect_TID_1500>
19. Fedorov, A., Clunie, D., Ulrich, E., Bauer, C., Wahle, A., Brown, B., Onken, M., Riesmeier, J., Pieper, S., Kikinis, R., Buatti, J. & Beichel, R. R. DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research. *PeerJ* **4,** e2057 (2016).
20. Clunie, D. *DICOM Structured Reporting*. (PixelMed Publishing, 2000). at <http://books.google.com/books?id=EVjOolUJNGUC&lpg=PP1&pg=PA6#v=onepage&q&f=false>
21. Moore, S. M., Maffitt, D. R., Smith, K. E., Kirby, J. S., Clark, K. W., Freymann, J. B., Vendt, B. A., Tarbox, L. R. & Prior, F. W. De-identification of Medical Images with Retention of Scientific Research Value. *Radiographics* **35,** 727–735 (2015).
22. Sharp, G. C., Li, R., Wolfgang, J. & Chen, G. Plastimatch: an open source software suite for radiotherapy image processing. *Proceedings of the* (2010). at <https://www.researchgate.net/profile/Maria_Spadea/publication/268523129_PLASTIMATCH-_AN_OPEN_SOURCE_SOFTWARE_SUITE_FOR_RADIOTHERAPY_IMAGE_PROCESSING/links/59de01a60f7e9bec3bae08ed/PLASTIMATCH-AN-OPEN-SOURCE-SOFTWARE-SUITE-FOR-RADIOTHERAPY-IMAGE-PROCESSING.pdf>
23. Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W.-L. & Wright, L. W. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* **40,** 30–43 (2007).
24. Langlotz, C. P. RadLex: a new method for indexing online educational materials. *Radiographics* **26,** 1595–1597 (2006).
25. Cornet, R. & de Keizer, N. Forty years of SNOMED: a literature review. *BMC Med. Inform. Decis. Mak.* **8 Suppl 1,** S2 (2008).
26. Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G. & Musen, M. A. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* **37,** W170–3 (2009).
27. Côté, R. G., Jones, P., Apweiler, R. & Hermjakob, H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* **7,** 97 (2006).
28. Zwanenburg, A., Leger, S., Vallières, M., Löck, S. & for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv [cs.CV]* (2016). at <http://arxiv.org/abs/1612.07003>
29. Opulencia, P., Channin, D. S., Raicu, D. S. & Furst, J. D. Mapping LIDC, RadLex™, and lung nodule image features. *J. Digit. Imaging* **24,** 256–270 (2011).
30. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J. C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S. & Kikinis, R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **30,** 1323–1341 (2012).
31. Eichelberg, M., Riesmeier, J., Wilkens, T., Hewett, A. J., Barth, A. & Jensch, P. Ten years of medical imaging standardization and prototypical implementation: the DICOM standard and the OFFIS DICOM toolkit (DCMTK). in *Proc. SPIE 5371, Medical Imaging 2004: PACS and Imaging Informatics* 57–68 (International Society for Optics and Photonics, 2004). doi:10.1117/12.534853

Data Citation

1. Andrey Fedorov, Matthew Hancock, David Clunie, Mathias Brockhausen, Jonathan Bona, Justin Kirby, John Freymann, Steve Pieper, Hugo Aerts, Ron Kikinis, Fred Prior. Standardized representation of the LIDC annotations using DICOM. (2018) The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2018.h7umfurq