LIDC-IDRI

Summary

The Lung Image Database Consortium image collection (LIDC-IDRI) consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. It is a web-accessible international resource for development, training, and evaluation of computer-assisted diagnostic (CAD) methods for lung cancer detection and diagnosis. Initiated by the National Cancer Institute (NCI), further advanced by the Foundation for the National Institutes of Health (FNIH), and accompanied by the Food and Drug Administration (FDA) through active participation, this public-private partnership demonstrates the success of a consortium founded on a consensus-based process.

Seven academic centers and eight medical imaging companies collaborated to create this data set which contains 1018 cases. Each subject includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists. In the initial blinded-read phase, each radiologist independently reviewed each CT scan and marked lesions belonging to one of three categories ("nodule > or =3 mm," "nodule <3 mm," and "non-nodule > or =3 mm"). In the subsequent unblinded-read phase, each radiologist independently reviewed their own marks along with the anonymized marks of the three other radiologists to render a final opinion. The goal of this process was to identify as completely as possible all lung nodules in each CT scan without requiring forced consensus.

Note: The TCIA team strongly encourages users to review pylidc and the DICOM representation of the annotations /segmentations included in this dataset before developing custom tools to analyze the XML version.

Data Access

Data Access

Click the **Download** button to save a ".tcia" manifest file to your computer, which you must open with the NBIA Data Retriever. Click the **Search** button to open our Data Portal, where you can browse the data collection and/or download a subset of its contents.

Data Type	Download all or Query/Filter
Images (DICOM, 125GB)	Download Search
DICOM Metadata Digest (CSV)	Download
Radiologist Annotations/Segmentations (XML format) (Note: see pylidc for assistance using these data)	Download
Nodule Size List (web)	Search
Nodule Counts by Patient (XLS)	Download
Patient Diagnoses (XLS)	Download

Click the Versions tab for more info about data releases.

Third Party Analyses of this Dataset

TCIA encourages the community to publish your analyses of our datasets. Below is a list of such third party analyses published using this Collection:

- Standardized representation of the TCIA LIDC-IDRI annotations using DICOM
- QIN multi-site collection of Lung CT data with Nodule Segmentations
- Segmentation of Pulmonary Nodules in Computed Tomography Using a Regression Neural Network Approach and its Application to the Lung Image Database Consortium and Image Database Resource Initiative Dataset
- Image Data Used in the Simulations of "The Role of Image Compression Standards in Medical Imaging: Current Status and Future Trends"

Detailed Description

Detailed Description

Collection Statistics	updated 3/21/2012
Modalities	CT (computed tomography) DX (digital radiography) CR (computed radiography) SEG (DICOM segmentations)*
Number of Patients	1010*
Number of Studies	1308*
Number of Series	1398*
Number of Images	244,617*
Image Size (GB)	125*

*__2/24/2020 Maintenance notes__: Corrected table entries only (no additional data). Added SEG to Modalities (previously present but not listed). Corrected number of subjects from 1006 to 1010, corrected number of studies from 1296 to 1308, corrected number of series from 1296 to 1398, corrected number of images from 243,185 to 244,617, corrected image size from 124 to 125. The corrections are to the table information only.

Reader Annotation and Markup

These links help describe how to use the .XML annotation files which are packaged along with the images in The Cancer Imaging Archive. The option to include annotation files in the download is enabled by default, so the XML described here will be included when downloading the LIDC-IDRI images unless you specifically uncheck this option. If you are only interested in the XML files or you have already downloaded the images you can obtain them here:

- LIDC-XML-only.zip

The following documentation explains the format and other relevant information about the XML annotation and markup files:

- XML File Documentation
- XML Base Schema - This file is called "voi array.xsd", and is central in defining tumors greater than or equal 3 mm in the datasets as well as defining the loci of non-nodules.
- Annotated XML File
- LIDC Radiologist Instructions for Spatial Location and Extent Estimates

Annotation and Markup Issues/Comments

1. For a subset of approximately 100 cases from among the initial 399 cases released, inconsistent rating systems were used among the 5 sites with regard to the spiculation and lobulation characteristics of lesions identified as nodules > 3 mm. The XML nodule characteristics data as it exists for some cases will be impacted by this error. We apologize for any inconvenience.
2. Also note that the XML files do not store radiologist annotations in a manner that allows for a comparison of individual radiologist reads across cases (i.e., the first reader recorded in the XML file of one CT scan will not necessarily be the same radiologist as the first reader recorded in the XML file of another CT scan).
3. March 2010: Contrary to previous documentation, the correct ordering for the subjective nodule lobulation and nodule spiculation rating scales stored in the XML files is 1=none to 5=marked. The issue of consistency noted above still remains to be corrected.
4. **On 2012-03-21 the XML associated with patient LIDC-IDRI-0101 was updated with a corrected version of the file.**
5. **Per May 2018, Please note that errors exist for two xml files, 044.xml and 191.xml, where one reader recorded one nodule as a "nodule >= 3 mm" but neglected to assign ratings for the nodule characteristics. On June 28, 2018 the files were updated with an explanation at the point of the error in the XML files.**
6. Subject LIDC-IDRI-0396 (139.xml) had an incorrect SOP Instance UID for position 1420. This was fixed on June 28, 2018.
7. Subject LIDC-IDRI-0510 has an assigned value of 5 for the internalStructure attribute in 187/255.xml. There is no 5th category for internalStructure so this should be considered invalid.

Nodule-Specific Details

- Nodule size list for the LIDC public cases - This link provides a list of available cases and the associated size of each identified nodule.
- lidc-idri nodule counts (6-23-2015).xlsx - This link provides an accounting of the total number of nodules for each LIDC-IDRI patient.

Diagnosis Data

For a limited set of cases, LIDC sites were able to identify diagnostic data associated with the case.

- tcia-diagnosis-data-2012-04-20.xls
- **Note:** This project has concluded and we are not able to obtain any additional diagnosis data beyond what is available in the above link.

Data was collected for as many cases as possible and is associated at two levels:

1. Diagnosis at the patient level (diagnosis is associated with the patient)
2. Diagnosis at the nodule level (where possible)

At each level, data was provided as to whether the nodule was:

1. Unknown (no data is available)
2. Benign or non-malignant disease
3. A malignancy that is a primary lung cancer
4. A metastatic lesion that is associated with an extra-thoracic primary malignancy

For each lesion, there is also information provided as to how the diagnosis was established including options such as:

1. unknown - not clear how diagnosis was established
2. review of radiological images to show 2 years of stable nodule
3. biopsy
4. surgical resection
5. progression or response

Software

pylidc

pylidc is an Object-relational mapping (using SQLAlchemy) for the data provided in the LIDC dataset. Some of the capabilities of pylidc include query of LIDC annotations in SQL-like fashion, conversion of the nodule segmentation contours into voxel labels, and visualization of segmentations as image overlays. If you find this tool useful in your research please cite the following paper:

> **Citation**
>
> Matthew C. Hancock, Jerry F. Magnan. **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**. *SPIE Journal of Medical Imaging*. Dec. 2016. http://dx.doi.org/10.1117/1.JMI.3.4.044504

MAX

MAX ("multi-purpose application for XML") performs nodule matching and pmap generation based on the XML files provided with the LIDC/IDRI Database. It also performs certain QA and QC tasks and other XML-related tasks.

MAX is written in Perl and was developed under RedHat Linux. It has been run under Windows.

Downloading MAX and its associated files implies acceptance of the following notice (also available here and in the distro as a text file):

```
Copyright 2006 - 2010
THE REGENTS OF THE UNIVERSITY OF MICHIGAN
ALL RIGHTS RESERVED

The software and supporting documentation was developed by the

         Digital Image Processing Laboratory
         Department of Radiology
         University of Michigan
         1500 East Medical Center Dr.
         Ann Arbor, MI 48109

It is funded in part by DHHS/NIH/NCI 1 U01 CA91099-01.

IT IS THE RESPONSIBILITY OF THE USER TO CONFIGURE AND/OR MODIFY THE SOFTWARE TO PERFORM THE
OPERATIONS THAT ARE REQUIRED BY THE USER.

THIS SOFTWARE IS PROVIDED AS IS, WITHOUT REPRESENTATION FROM THE UNIVERSITY OF MICHIGAN AS
TO ITS FITNESS FOR ANY PURPOSE, AND WITHOUT WARRANTY BY THE UNIVERSITY OF MICHIGAN OF ANY
KIND, EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OF
MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE REGENTS OF THE UNIVERSITY OF
MICHIGAN SHALL NOT BE LIABLE FOR ANY DAMAGES, INCLUDING SPECIAL, INDIRECT, INCIDENTAL,
OR CONSEQUENTIAL DAMAGES, WITH RESPECT TO ANY CLAIM ARISING OUT OF OR IN CONNECTION WITH
THE USE OF THE SOFTWARE, EVEN IF IT HAS BEEN OR IS HEREAFTER ADVISED OF THE POSSIBILITY
OF SUCH DAMAGES.

PERMISSION IS GRANTED TO USE, COPY, CREATE DERIVATIVE WORKS AND REDISTRIBUTE THIS SOFTWARE
AND SUCH DERIVATIVE WORKS FOR ANY PURPOSE, SO LONG AS THIS ENTIRE COPYRIGHT NOTICE,
INCLUDING THE GRANT OF PERMISSION, AND DISCLAIMERS, APPEAR IN ALL COPIES MADE; AND SO LONG
AS THE NAME OF THE UNIVERSITY OF MICHIGAN IS NOT USED IN ANY ADVERTISING OR PUBLICITY
PERTAINING TO THE USE OR DISTRIBUTION OF THIS SOFTWARE WITHOUT SPECIFIC, WRITTEN PRIOR
AUTHORIZATION.
```

DISCLAIMER: MAX is not guaranteed to process all input correctly. Possible errors include (but are not limited to) the inability to process correctly some types of nodule ambiguity (where nodule ambiguity refers to overlap between nodule markings having complicated shapes or to overlap between a nodule marking and a non-nodule mark).

Download the distro (max-V107.tgz); view/download ReadMe.txt (a text file that is also included in the distro).

LIDC 2 Image Toolbox (Matlab)

This tool is a community contribution developed by Thomas Lampert. It is designed for extracting individual annotations from the XML files and converting them, and the DICOM images, into TIF format for easier processing in Matlab (LIDC-IDRI dataset). It is available for download from: https://sites.google.com/site/tomalampert/code.

Citations & Data Usage Policy

Citations & Data Usage Policy

This collection is freely available to browse, download, and use for commercial, scientific and educational purposes as outlined in the Creative Commons Attribution 3.0 Unported License. See TCIA's Data Usage Policies and Restrictions for additional details. Questions may be directed to help@cancerimagingarchive.net.

Please be sure to include the following citations and attributions in your work if you use this data set:

Data Citation

Armato III, SG; McLennan, G; Bidaut, L; McNitt-Gray, MF; Meyer, CR; Reeves, AP; Zhao, B; Aberle, DR; Henschke, CI; Hoffman, Eric A; Kazerooni, EA; MacMahon, H; van Beek, EJR; Yankelevitz, D; Biancardi, AM; Bland, PH; Brown, MS; Engelmann, RM; Laderach, GE; Max, D; Pais, RC; Qing, DPY; Roberts, RY; Smith, AR; Starkey, A; Batra, P; Caligiuri, P; Farooqi, Ali; Gladish, GW; Jude, CM; Munden, RF; Petkovska, I; Quint, LE; Schwartz, LH; Sundaram, B; Dodd, LE; Fenimore, C; Gur, D; Petrick, N; Freymann, J; Kirby, J; Hughes, B; Casteele, AV; Gupte, S; Sallam, M; Heath, MD; Kuhn, MH; Dharaiya, E; Burns, R; Fryd, DS; Salganicoff, M; Anand, V; Shreter, U; Vastagh, S; Croft, BY; Clarke, LP. (2015). **Data From LIDC-IDRI**. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX

Publication Citation

Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Van Beeke EJ, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DP, Roberts RY, Smith AR, Starkey A, Batrah P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Casteele AV, Gupte S, Sallamm M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY. **The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans**. Medical Physics, 38: 915--931, 2011. DOI: https://doi.org/10.1118/1.3528204

TCIA Citation

Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. (2013) **The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository**, Journal of Digital Imaging, Volume 26, Number 6, pp 1045-1057. DOI: https://doi.org/10.1007/s10278-013-9622-7

In addition, please be sure to include the following attribution in any publications or grant applications along with references to appropriate LIDC publications:

> *The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.*

Other Publications Using This Data

See the LIDC-IDRI section on our Publications page for other work leveraging this collection. If you have a publication you'd like to add please contact the TCIA Helpdesk.

Versions

Version 3 (Current): Updated 2015/07/27

Data Type	Download all or Query/Filter
Images (DICOM, 125GB)*	Download Search (Requires the NBIA Data Retriever.)
DICOM Metadata Digest (CSV)	Download
Radiologist Annotations/Segmentations (XML)	Download
Nodule Size List (web)	Search
Nodule Counts by Patient (XLS)	Download
Patient Diagnoses (XLS)	Download

*Replace any manifests downloaded prior to 2/24/2020. Please download a new manifest by clicking on the download button in the *Images* row of the table above. Manifests downloaded prior to 2/24/2020 may not include all series the collection.

Prior to 7/27/2015, many of the series in the LIDC-IDRI collection,had inconsistent values in the DICOM Frame of Reference UID, DICOM tag (0020,0052). Each image had a unique value for Frame of Reference (which should be consistent across a series). This has been corrected. In addition, the following tags, which were present (but should not have been), were removed: (0020,0200) Synchronization Frame of Reference, (3006,0024) Referenced Frame of Reference, and (3006,00c2) Related Frame of Reference.

Version 2: Updated 2012/03/21

On 2012-03-21 the XML associated with patient LIDC-IDRI-0101 was updated with a corrected version of the file. The old version is still available if needed for audit purposes.

Version 1:

There was a "pilot release" of 399 cases of the LIDC CT data via the NCI CBIIT installation of NBIA. The LIDC-IDRI collection contained on TCIA is the complete data set of all 1,010 patients which includes all 399 pilot CT cases plus the additional 611 patient CTs and all 290 corresponding chest x-rays. A table which allows mapping between the old NBIA IDs and new TCIA IDs can be downloaded for those who have obtained and analyzed the older data.

For a subset of approximately 100 cases from among the initial 399 cases released, inconsistent rating systems were used among the 5 sites with regard to the spiculation and lobulation characteristics of lesions identified as nodules > 3 mm. The XML nodule characteristics data as it exists for some cases will be impacted by this error. We apologize for any inconvenience.

Contrary to previous documentation (prior to March 2010), the correct ordering for the subjective nodule lobulation and nodule spiculation rating scales stored in the XML files is 1=none to 5=marked. The issue of consistency noted above still remains to be corrected.