An Analysis of Early Studies Released by the Lung Imaging Database Consortium (LIDC)

Wesley D. Turner, Timothy P. Kelliher, James C. Ross, and James V. Miller*

GE Research Center, Niskayuna NY 12309, USA

Abstract. Lung cancer remains an ongoing problem resulting in substantial deaths in the United States and the world. Within the United states, cancer of the lung and bronchus are the leading causes of fatal malignancy and make up 32% of the cancer deaths among men and 25% of the cancer deaths among women. Five year survival is low, (14%), but recent studies are beginning to provide some hope that we can increase survivability of lung cancer provided that the cancer is caught and treated in early stages. These results motivate revisiting the concept of lung cancer screening using thin slice multidetector computed tomography (MDCT) protocols and automated detection algorithms to facilitate early detection. In this environment, resources to aid Computer Aided Detection (CAD) researchers to rapidly develop and harden detection and diagnostic algorithms may have a significant impact on world health. The National Cancer Institute (NCI) formed the Lung Imaging Database Consortium (LIDC) to establish a resource for detecting, sizing, and characterizing lung nodules. This resource consists of multiple CT chest exams containing lung nodules that seveal radiologists manually countoured and characterized. Consensus on the location of the nodule boundaries, or even on the existence of a nodule at a particular location in the lung was not enforced, and each contour is considered a possible nodule. The researcher is encouraged to develop measures of ground truth to reconcile the multiple radiologist marks. This paper analyzes these marks to determine radiologist agreement and to apply statistical tools to the generation of a nodule ground truth. Features of the resulting consensus and individual markings are analyzed.

1 Introduction

Despite warnings stretching back 40 years – the initial United States Surgeon General's Report on Smoking was issued in 1964 and warning labels have been required on cigarettes sold in the United States since 1969 – lung cancer remains a serious health threat in the world with an estimated 1 million deaths in 2000 [1]. For the United States, lung cancer is the single largest fatal malignancy and is second only to heart disease in yearly fatalities. There are few effective treatments and five year mortality is approximately 14% [2,3] due largely to advanced stage of the disease at diagnosis. Despite this, lung cancer screening is not recommended, even for at risk populations, based largely

* This publication was supported by the DOD and the Medical University of South Carolina under DOD Grant No. W81XWH-05-1-0378. Its contents are soley the responsibility of the authors and do not necessarily represent the official views of the Department of Defense or the Medical University of South Carolina.

R. Larsen, M. Nielsen, and J. Sporring (Eds.): MICCAI 2006, LNCS 4191, pp. 487–494, 2006.

on statistically powerful studies carried out in the 1970s that showed little benefit to lung cancer screening. However, imaging advances over recent years – most notably the advent of thin slice MDCT – have prompted revisiting the notion of screening.

Recent studies of stage I cancers demonstrate that the diagnosis of lung cancer prior to metasteses can give five year survival rates as high as 70% [4,1] and other studies [5] equate small, subtle cancers with early stage disease. But before screening becomes widely accepted, issues involving overdiagnosis, radiation dose, and appropriate population selection must be addressed. However, a dose-limiting screening protocol that enables early nodule detection while limiting overdiagnosis appears within reach. One of the promising vehicles for early detection is the low dose, Multi-Detector Computed Tomography (MDCT) scan. MDCT has been shown to be an effective method for detecting small ($<$1cm) nodules in the lungs that may be cancerous or pre-cancerous lesions. However, the rigors of reading a large number of MDCT chest scans with upwards of 500 images per scan, combined with the subtlety of some of nodules of interest argue for computer aided detection (CAD) as a necessity for a rational screening program.

In this environment, the National Cancer Institute (NCI) formed the Lung Imaging Database Consortium (LIDC) to study the barriers to effective CAD development and to develop a database as a national resource that can be used to expedite development [6]. One of the major barriers to the development and evaluation of effective CAD devices is the absence of rigorous ground truth (GT). Unlike other cancer screening applications, the determination of malignancy in a discovered lung lesion is not undertaken for small, subtle lesions due to the likelihood of morbidity during the procedure. Although histological analysis of biopsy samples provides the most reliable assesment of malignancy, radiological diagnosis based on medical images currently serves as a surrogate for an actual malignancy determination. Performance of CAD algorithms tends to be very sensitive to the choice of GT. Consider the chart in Figure 1, which represents a set of lung scans read by three radiologists. The data was generated based on lung cancer screening exams in the GE Global Research cancer database. The y-axis shows how many nodules were detected by one, two or all three radiologists reading the exams. Depending on whether single radiologist GT, majority GT, or unanimous GT is chosen, the performance of a typical CAD system will change drastically. As an example, a CAD algorithm with 100% sensitivity on unanimous GT could have much lower sensitivity on single reader GT, and an algorithm with 100% specificity for single reader GT could gain hundreds of false positives using unanimous GT.

To get the best possible GT under these constraints, and to capture nodule characteristics such as spiculations, lobulations and density, the LIDC asked multiple radiologists to perform a series of blinded and unblinded reads on a number of CT lung images. Each radiologist first read the lung cases without knowledge of how his/her colleagues marked the exams and placed contours around the periphery of nodules found. The radiologists also described the nodules with the features in Table 1. Once all initial reads where completed, a further round of unblinded reads where performed where the radiologists could modify their own contours based on the contours of their colleagues. In the end, all four sets of final contours were saved and provided as annotations to the exams. This allows CAD and cancer researchers to determine an appropriate method to

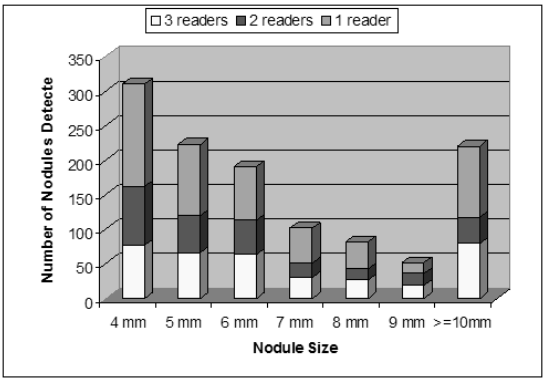

Fig. 1. The variability of nodule ground truth based on radiological consensus

Table 1. Nodule characeristics captured in the LIDC datasets

Characteristic	Definition				
Subtlety	1 - Extremely	2 - Moderately	3 - Fairly	4 - Moderately	5 - Obvious
Texture	1 - non/GGO	2	3 - Part/Mixed	4	5 - Solid
Margin	1 - Poorly Defined	2	3	4	5 - Sharp
Sphericity	1 - Linear	2	3 - Ovoid	4	5 - Round
Spiculation	1 - Marked	2	3	4	5 - None
Lobulation	1 - Marked	2	3	4	5 - None

combine the GT for their research so as to maximize detection and diagnostic capabilities of their algorithms.

In this paper, we investigate the radiologist GT in the first twenty-nine LIDC datasets released. We use statistical tools such as Simultaneous Truth and Performance Level Estimation (STAPLE) [7] available in the Insight Segmentation and Registration ToolKit (ITK) [8] to both extract estimations of GT contours from the multiple radiologist marks and to characterize reader performance.

2 Methods

The LIDC in cooperation with the NCI are collecting and distributing a database of lung MDCT scans with annotations of nodules and nodule charateristics. The public database currently stands at 30 datasets, of which we were able to process 29. The dataset size is growing and will reach 400 cases before the end of the project. We downloaded the cases from the LIDC public ftp site [9] (the data is also available at [10]). This data is stored in anonymized DICOM format that eliminates patient identification elements from the images, but preserves other data annotations such as dosage and reconstruction parameters as well as the make and model of the scanner for each study. The image data is accompanied by an XML file with information from the experts' reads. The XML details nodules smaller than 3mm, nodules larger than 3mm and non-nodules. In this analysis we have focused on nodules larger than 3mm.

Using ITK we read each image series and the corresponding XML data, which gives a list of points defining a 2D contour at a specific Z-location within the CT image stack. We convert these contours to label images in which only the interior voxels are labelled. This is in keeping with the instructions given to the radiologists, who were instructed to mark the nodules one pixel outside the border.

The next step is to ensure that the same nodule is labeled consistently in each of the reader maps. This is done by comparing the bounding boxes of the nodules. If the centroid of a bounding box overlaps with the bounding box of a nodule by a different radiologist, we consider it to represent the same nodule. After the nodules are cross-compared each resulting unique nodule is assigned a label and the original label maps are adjusted to use the new labeling.

The result of this processing is a set of volumes labeled with consistent nodule numberings across all of the readers for each series from the database. We write these volumes back to disk as DICOM series. While this is not the most compact storage choice, it has the advantage that all of the original DICOM information is carried with the label maps, minimizing the opportunity for error later in the process and allowing the nodule contours to be viewed with standard DICOM viewing tools. The result is a set D of label maps where $D_{m,n}$ is the set of voxels from reader m with label n.

2.1 Calculating Ground Truth

Given the consistently labeled nodule volumes representing a set of expert opinions, we combine the individual information into a common ground truth. There are a number of ways in which this can be done. We compute several and then contrast the results. The most expansive combination that we compute for a given nodule with label n is the union of all reads.

$$T_{max_n} = \bigcup A,\ A \in D_{m,n}\ \forall m. \tag{1}$$

For this case any voxel labeled as part of nodule n by any of the readers is considered to be a valid part of n. This is equivalent to saying that all of the readers have perfect specificity, while some may have reduced sensitivity. That is all marked voxels are valid parts of the nodules and no marks represent false positive voxels.

Conversly, we also compute the smallest estimate of ground truth by only considering those voxels that were selected by the entire set of readers.

$$T_{min_n} = \bigcap A,\ A \in D_{m,n}\ \forall m. \tag{2}$$

In this case, there are no false negative voxels; readers are accorded perfect sensitivity while potentially suffering from reduced specificity.

Together T_{max} and T_{min} form the bounds on that which we will consider to be truth. For some questions these two estimates may be sufficient representations of the truth. Questions concerning whether the case needs further consideration may rely soly on the T_{max} estimate. Other questions that look at the reproducibility of results may consider $\|T_{min}\|/\|T_{max}\|$. As this ratio approaches 1 the agreement on nodule definition is approaching consensus. The size and shape variation of nodules that fall within these

bounds, however, is too wide to use in comparing algorithms that attempt to segment nodules for the purposes of computing measures of nodule features such as lobulation and spiculation.

To arrive at a more accurate estimate of ground truth, suitable for use in scoring segmentation algorithms, we must consider truth not as just a binary decision, as in (1) and (2), but as a continuum from which we will select. To arrive at this narrower ground truth, we start by marking all the voxels which are in the complement of T_{max} as having zero probability of being a part of the truth and at the same time mark the voxels that are elements of T_{min} as having a 100% probability of being part of the truth. This leaves $T_{max} - T_{min}$ as those voxels whose probability is uncertain. There are two methods in common use for selecting the probabilities to assign to these voxels.

For the first method consider each radiologist marking as a vote for including the voxel in the nodule [11] as below:

$$v_{n_{i,j,k}} = \sum_{m=1}^{R} V_{m,n}(i,j,k)/R, \tag{3}$$

$$V_{m,n}(i,j,k) = \begin{cases} 1 & I_{i,j,k} \in D_{m,n} \\ 0 & otherwise \end{cases} \tag{4}$$

where $v_{n_{i,j,k}}$ is a voxel in the $pmap$ for nodule n at image position (i,j,k), $I_{i,j,k}$ is the voxel in the label map, and R is the number of radiologists who provided annotations.

This voting method results in the creation of a $pmap$ representing the probability that a voxel is a part of a nodule. As presented, the voting method classifies all reader's opinions as equally valid. While it is possible to make a simple extension to weight readers differently in the voting based on their relative expertise, this has not been the practice in this domain.

The second method for computing probabilities for the uncertain voxels is STAPLE [7]. Staple simultaneously computes the consensus ground truth and the specificity and sensitivity for each of the readers. It employs an Expectation Maximization (EM) algorithm to jointly optimize these values. As with the voting method a $pmap$ is produced with values at 0 for voxels definitely outside the nodule and 1 for voxels definitely inside based on the available manual markings. Voxels where certainty cannot be achieved are marked in the interval $(0,1)$. Furthermore, STAPLE can be preconditioned with the expected sensitivities and specificities of the readers to guide the optimization.

Both the voting and STAPLE $pmaps$ represent a continuum of possible ground truths. They can be thresholded at a particular point to force a binary decision on what will be considered as truth for a particular calculation. In practice, the 50% level is often used as dividing line between regions in computer vision applications.

3 Results

Table 2 shows a comparison of ground truthing techniques for several representative cases from the LIDC database. In each column the total number of nodule-labeled voxels is given. It is evident by comparing the discrepancy in T_{min} and T_{max} values that intra-observer variability is an issue. As expected, the thresholded $pmaps$ produce totals that lie between the corresponding T_{min} and T_{max} values. The fact that the total

number of voxels remains unchanged as the $pmaps$ are thresholded at different levels indicates that probability values are tightly grouped either below the 25% level or above the 75% level.

Table 2. Comparison of Ground Truth Approaches

Case	T_{max}	T_{min}	$pmap_{25}$	$pmap_{50}$	$pmap_{75}$
1	244	88	160	160	160
2	596	175	276	276	276
3	3049	1100	1630	1630	1630
4	953	233	484	484	484
5	6325	2356	4092	4092	4092

Figure 2 shows models generated using the Visualization Toolkit's discrete marching cubes implementation [12,13]. Not only is inter-observer variability clear, but it is also evident that for some cases substantial step artifacts exist between slices.

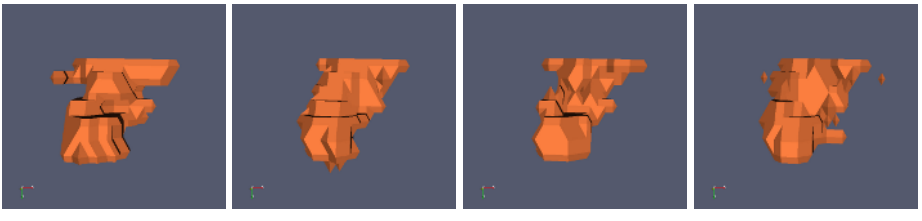

Fig. 2. Models of the same nodule segmented by different readers

Figure 3 shows reader segmentations for a representative nodule, as well as the corresponding $pmap$ generated by STAPLE. Variability across readers is clear. STAPLE computed sensitivity averages 66% over the 24 nodule positive cases. Specificity is nearly 100% (99.9993%) over the same cases, but this to be expected because the vast number of voxels in the cases are not marked as part of any nodule. STAPLE computed sensitivity on a per nodule basis is higher (82%) for the 26 unanimously detected nodules reflecting radiologist disagreement about the actual presence of some subtle nodules.

4 Conclusion

Diagnosis of lung nodules and recommendation of potentially costly follow-up depend on the accurate assesment of nodule characteristics. Inter- and intra-observer variability in such assesments indicates the complexity of the issue and motivates the development of CAD techniques to assist human readers. The LIDC database provides an excellent resource against which CAD algorithms can be scored. For accurate evaluation, however, a notion of ground truth must be derived from the hand segmentations provided

Fig. 3. From left to right: (a) zoomed CT region around a nodule, (b) reader 0 segmentation, (c) reader 1 segmentation, (d) reader 2 segmentation, (e) reader 3 segmentation, and (f) log(pmap) scaled to range from 0 to 255 in order to highlight probability drop-offs. The probabilities represented in this pmap are 100%, 99.9%, 6.1%, 4.7%, and 1.2%. Reader sensitivity and specificity for this nodule as determined by STAPLE are, respectively: (0.903, 0.999), (0.876, 1), (0.781, 1), (0.846, 1). Mean sensitivity and specificity are (0.851, 1).

by the expert radiologists. In this paper we have proposed and compared the use of several ground-truthing techniques: the union of all reads, the intersection of all reads, and STAPLE. We performed STAPLE analysis on a per-case basis to better leverage the detection variability of the cases, but better segmentation results might be obtained by running on a per-nodule basis. We currently plan to release $pmap$ data to a public ftp site for free download.

It would be interesting to correlate dosage and reconstruction settings with human segmentation results. Such an analysis on segmentation contours could be performed by investigating $pmaps$ produced by the STAPLE algorithm. Image acquisition parameters that result in $pmaps$ that are tightly grouped around 0% or 100% probability suggest a protocol that mitigates human variability in contour deliniation. As another next step we plan to develop automated 2D/3D nodule measures and correlate them with radiologist reported measures. Such measures would address the issue of human variability.

Hand segmentations performed on 2D slices can lead to "discontinuous" segmentations as viewed in 3D – no doubt as a function of the oftentimes amorphous nature of nodules as well as the slice thickness of CT images and partial voluming. In order to mitigate step-like artifacts, we propose the use of additional hand segmentation tools that would augment human contouring. One idea would be to provide a 3D model immediately after the complete segmentation of a nodule for evaluation. Performing segmentations in coronal and sagittal reconstructions in addition to axial could also prove beneficial as could segmentations on thinner slice protocols.

References

1. Humphrey, L., Teutsch, S., Johnson, M.: Lung cancer screening with sputum cytology examination, chest radiography, and computed tomography: An update for the U.S. preventive services task force. Annal Intern Med (2004) 740–753
2. Fry, W.A., Menck, H.R., Winchester, D.P.: The national database report on lung cancer. Cancer **77** (1996) 1947–1955
3. Dodd, L., Wagner, R., Armato III, S., McNitt-Gray, M.F., Beiden, S., Chan, H.P., Gur, D., McLennan, G., Metz, C., Petrick, N., Sahiner, B., Sayre, J., The Lung Imaging Database Consortium Group: Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: Contemporary research topics relevant to the lung image database consortium. Acad Radiol (2004) 462–475

4. Henschke, C.I., Naidich, D.P., Yankelevitz, D.F., McGuinness, G., McCauley, D.I., Smith, J.P., Libby, D., Pasmantier, M., Vazquez, M., Koizumi, J., Flieder, D., Altorki, N., Miettinen, O.S.: Early lung cancer action project: Initial findings on repeat screenings. Cancer **92** (2001) 153–159
5. Suzuki, K., Kusumoto, M., S., W., Tsuchiya, R., Asamura, H.: Radioloigic classification of small adenocarcinoma of the lung: Radiologic-pathologic correlation and its prognostic impact. Ann Thora Surg (2006) 413–420
6. Armato III, S., McClennan, G., McNitt-Gray, M.F., Meyer, C., Yankelevitz, D., Aberle, D., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Reeves, A.P., Croft, B.Y., Clarke, L.P., For the Lung Image Database Consortium Research Group: Lung image database consortium: Developing a resource for the medical imaging research community. In: Radiology, RSNA (2004) 739–748
7. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23** (2004) 903–921
8. Ibáñez, L., Schroeder, W., Ng, L., Cates, J., The Insight Software Consortium: The ITK Software Guide Second Edition. http://www.itk.org (2005)
9. NCI/LIDC: ftp://ncicbfalcon.nci.nih.gov/lidc/lidc_release_0001/ (2006)
10. NCI/LIDC: http://ncia.nci.nih.gov (2006)
11. NCI/LIDC: http://imaging.cancer.gov/reportsandpublications/reportsandpresentations/firstdataset (2006)
12. Lorensen, W., Miller, J.V., Padfield, D., Ross, J.C.: Creating models from segmented medical images. In: Medicine Meets Virtual Reality. (2005)
13. Schroeder, W., Martin, K., Lorensen, W.: The Visualization Toolkit. Kitware Inc. (2004)