Comparison of computer versus manual determination of pulmonary nodule volumes in CT scans

Alberto M. Biancardi[a], Anthony P. Reeves[a], Artit C. Jirapatnakul[a], Tatiyana Apanasovitch[b], David Yankelevitz[c], and Claudia I. Henschke[c]

[a] School of Electrical and Computer Engineering, Cornell University, Ithaca, NY
[b] School of Operations Research and Information Engineering, Cornell University, Ithaca, NY
[b] Department of Radiology, Weil Medical College of Cornell University, New York, NY

ABSTRACT

Accurate nodule volume estimation is necessary in order to estimate the clinically relevant growth rate or change in size over time. An automated nodule volume-measuring algorithm was applied to a set of pulmonary nodules that were documented by the Lung Image Database Consortium (LIDC). The LIDC process model specifies that each scan is assessed by four experienced thoracic radiologists and that boundaries are to be marked around the visible extent of the nodules for nodules 3 mm and larger. Nodules were selected from the LIDC database with the following inclusion criteria: (a) they must have a solid component on a minimum of three CT image slices and (b) they must be marked by all four LIDC radiologists. A total of 113 nodules met the selection criterion with diameters ranging from 3.59 mm to 32.68 mm (mean 9.37 mm, median 7.67 mm). The centroid of each marked nodule was used as the seed point for the automated algorithm. 95 nodules (84.1%) were correctly segmented, but one was considered not meeting the first selection criterion by the automated method; for the remaining ones, eight (7.1%) were structurally too complex or extensively attached and 10 (8.8%) were considered not properly segmented after a simple visual inspection by a radiologist. Since the LIDC specifications, as aforementioned, instruct radiologists to include both solid and sub-solid parts, the automated method core capability of segmenting solid tissues was augmented to take into account also the nodule sub-solid parts. We ranked the distances of the automated method estimates and the radiologist-based estimates from the median of the radiologist-based values. The automated method was in 76.6% of the cases closer to the median than at least one of the values derived from the manual markings, which is a sign of a very good agreement with the radiologists' markings.

Keywords: Methods: quantitative image analysis, Modalities: X-ray CT, Diagnostic Task: diagnosis, Diagnostic Task: response to therapy, volumetric nodule measurement.

1. INTRODUCTION

For lung nodules, estimation of growth rates or of size changes plays a fundamental role both in clinical practice and in pharmacological research because it enables the determination of the probability of a nodule malignancy or of the efficacy of a therapy. The accuracy and precision of those estimations are linked, in turn, to the accuracy and precision of the absolute volume estimations performed on the single imaged instances. With the current trend toward higher and higher resolutions on the axial dimension, manual volumetric measurement is becoming more and more demanding, being both time intensive and subject to fatigue; additionally it has been shown[1, 2] to have a high intra- and inter- observer variability, albeit better than mono- and bi-dimensional measures.

A reliable automated algorithm would require much less time, requiring only a quality-control review, and would essentially eliminate the problem of variability by applying the same set of rules to each of the sequential scans: this is why several efforts have been actively developed.[3–8] The difficulty of this approach is now shifted toward the need to calibrate and validate such methods and, currently, the only accepted source for the definition of a gold standard is based on nodule boundary markings performed by expert radiologists.

Send correspondence to A. M. Biancardi e-mail: amb284@cornell.edu, phone: +1.607.254.8819, fax: +1.607.255.9072

The Lung Image Database Consortium[9] is a cooperative program, started by the National Institutes of Health in 2000, aiming at the creation of a large database of documented whole lung CT scans for the development and the evaluation of different CAD approaches. The LIDC process model[10, 11] asks radiologists to draw the visible boundary of any nodule estimated to be larger than 3 mm and does not require a strict consensus, allowing the presence of multiple boundaries whenever more than one radiologist considered an opacity eligible for the database. The availability of such annotated nodules makes the LIDC database an invaluable source for the definition of a ground truth against which automated methods can be tested. This paper describes the validation that was carried out by comparing the volumetric sizes computed by our research system with those derived from the LIDC archive.

In this paper an automated method under development by our research group was first tested in its standard version capable of segmenting the solid tissue of nodules. However, as this criterion is much more stringent than what specified by the LIDC for the manual markings, it does not take into account the sub-solid components that are marked by the radiologists. Therefore, the automated method core capability of segmenting solid tissues was augmented to take into account the nodule sub-solid parts in order to better simulate the LIDC specifications.

2. MATERIALS AND METHODS

The comparison was performed on whole-lung CT scans provided by the LIDC archive. The LIDC process model specifies that each scan is assessed by four experienced thoracic radiologists and that, for nodules three mm and larger, boundaries are to be marked, in every axial image in which they appear, around the visible extent of the nodules, which includes the whole range of radiologically detectable tissues from sub-solid to solid. Radiologist may also mark inner boundaries to express the fact that a portion inside the outer boundary does not belong to the actual nodule. The nodule boundaries were processed to determine the nodule centroid and four values for its volume, one for each of the radiologists' markings. The total lesion volume is estimated by counting the number of nodule pixels in each of the image slices and then multiplying their sum by the voxel volume;[12] this method is frequently used in CAD tools. Pixels belonging to any excluded inner regions do not belong to the nodule region and therefore are not counted were computing the nodule volume. The centroids were used as the seed points for the automated algorithm. Nodules were selected from the LIDC database with the following inclusion criteria: (a) they must have a solid component on a minimum of three CT image slices and (b) they must be marked by all four LIDC radiologists.

For this paper 265 whole-lung CT scans were available, of which 197 had nodules documented with radiologists' boundaries. All of the 197 scans were acquired from multi-detector row CT scanners with pixel size ranging from 0.508 to 0.946 mm (average 0.66 mm) and an axial slice thickness ranging from 0.625 to 3.000 mm (average 1.7 mm, median 1.8 mm). The tube current ranged from 40 to 582 mA (average 177.5 mA, median 160 mA), tube voltage range was for more than half of the cases 120kVp with the remaining ones having voltages equal to 130kVp (8), 135kVp (23), and 140 kVp (28).

A set of three-dimensional regions was obtained by executing our segmentation algorithm,[8, 13] which can be summarized in the following steps. After determining a sufficiently ample bounding box containing the nodule, the raw scan images are resampled to an isotropic space by tri-linear interpolation where all the actual segmentation will take place. After performing a grey level thresholding to extract solid component, the data is analyzed to find out to which of the four standard models, (well-circumscribed, vascularized, pleural tail, or juxtapleural) the nodule is closest. If the nodule has pleural tail or is juxtapleural, a cut surface is determined to shrink the nodule bounding box and remove the attachment. Then a candidate solid part is extracted by considering all the connected components with voxels greater than or equal to a given intensity threshold. Except for well-circumscribed nodule, vessel structures, having a CT attenuation similar to the nodule one, will be included in the candidate solid component; the application of iterative morphological filters allows the method to remove those structures. Additionally, by processing data from the original bounding box the sub-solid component is extracted in a similar way as the solid part, but using a different threshold value. The final stage of the nodule segmentation process is a quick expert review to detect major segmentation errors. We distinguish between measurement errors, in which the precise location of the nodule boundary is at issue, and catastrophic errors, in which an incorrect region is segmented.

Automated segmentation methods use heuristics to separate attached structures from solid nodules that do not always provide the correct outcome. Over-segmentation failures occurs when an attached structure such as a vessel or the chest wall is considered to be part of the lesion. Under-segmentation occurs typically for lesion with complex shapes when a section of the lesion is not identified as part of the lesion. Catastrophic errors are usually immediately recognized by inspection of an experienced user. One possibility is to have the user correct this error by manual intervention. In this study we rejected cases where the automatic algorithm had such a failure.

For the first version of the automated method only the volumes computed for the nodule solid component were used. The second version, on the other hand, used a linear combination of the estimated volumes for both the solid and the sub-solid parts. The two versions of the automated method were then compared to the volumes determined from the radiologists markings. The median of the radiologist-based volumes was computed and then the distances, from this median, of the automated method estimates and the radiologist-based estimates were ranked. We considered measures that were not the furthest from the median (i.e. was not ranked fifth) to be in agreement.

3. RESULTS

Within the full set of 265 scans, 197 had nodules with marked boundaries and 128 nodules met the second selection criterion, being marked by all the four radiologists. Of these, 113 had a solid component on a minimum of three CT image slices, according to the radiologists' markings. The size range of the selected nodules, expressed as the diameter length of an equivalent sphere having the same volume as the estimates, was from 3.59 mm to 32.68 mm (mean 9.37 mm, median 7.67 mm) as shown in Figure 1. A total of 95 nodules (92.9%) were correctly segmented, but one was considered by the automated method as not fulfilling requirement (a), having a solid component of less than 3 slices; for the remaining nodules, eight (7.1%) were structurally too complex or extensively attached, resulting in a properly managed termination of the automated procedure and ten (8.8%) were considered as having a major segmentation problem after the quick quality-control review by a radiologist.

Table 1 shows the rankings of the first and second version of the automated method when compared to the radiologists' estimates according to the absolute size of the nodules. Table 2 shows the rankings of both versions of the automated method according to the distances from the median of the radiologists' estimates.

Table 1. Rankings of volume estimates by the automated method when compared to the radiologists' estimates according to the absolute value. The numbers in bold indicate when there is an agreement of the automated method with the radiologists.

Rankings according to size

	1st	2nd	3rd	4th	5th
solid only	30	**26**	**19**	**8**	11
solid and non-solid	16	**25**	**21**	**19**	13

Table 2. Rankings of volume estimates by the automated method when compared to the radiologists' estimates according to the distance from the radiologist-estimate median. The numbers in bold indicate when there is an agreement of the automated method with the radiologists.

Rankings according to distance from median

	1st	2nd	3rd	4th	5th
solid only	**19**	**0**	**20**	**25**	30
solid and non-solid	**21**	**0**	**23**	**28**	22

4. DISCUSSION

A total of 94 nodules were analyzed for this study. Out of the 113 meeting the selection criteria, 1 was properly segmented, but considered too small by the automated method, 8 caused the segmentation program to signal

the unavailability of a result and 10 were considered having major problems at the final visual inspection; all of them were excluded from the statistical computations. Figure 2 shows an example were a vascular structure was not recognized and removed.

It is worth underlining that, per the LIDC specifications, the nodule outer boundary was chosen to be made of those pixels that were just outside the largest region that could be linked to the nodule presence. By this definition, all the sub-solid parts as well as the solid component should be included in the marked region. In some cases, pixels with minimal added opacity with respect to the surrounding parenchyma were marked as belonging to the nodule region.

A critical point in our study was the definition of an agreement criterion; i.e. the definition of a boolean test that, given the five volume estimates, would tell whether the automated method value could be considered in agreement with the radiologist-based values. A major source of difficulty lies in the anonymization not only of the scan DICOM data, but also of the respective annotations. This means that a number of key aspects cannot be known or taken for granted:

- which sites produced the markings (there are five sites cooperating to the LIDC, but only four readers[11]);
- whether a site has multiple readers and, therefore, the actual number of radiologists that contributed the annotations;

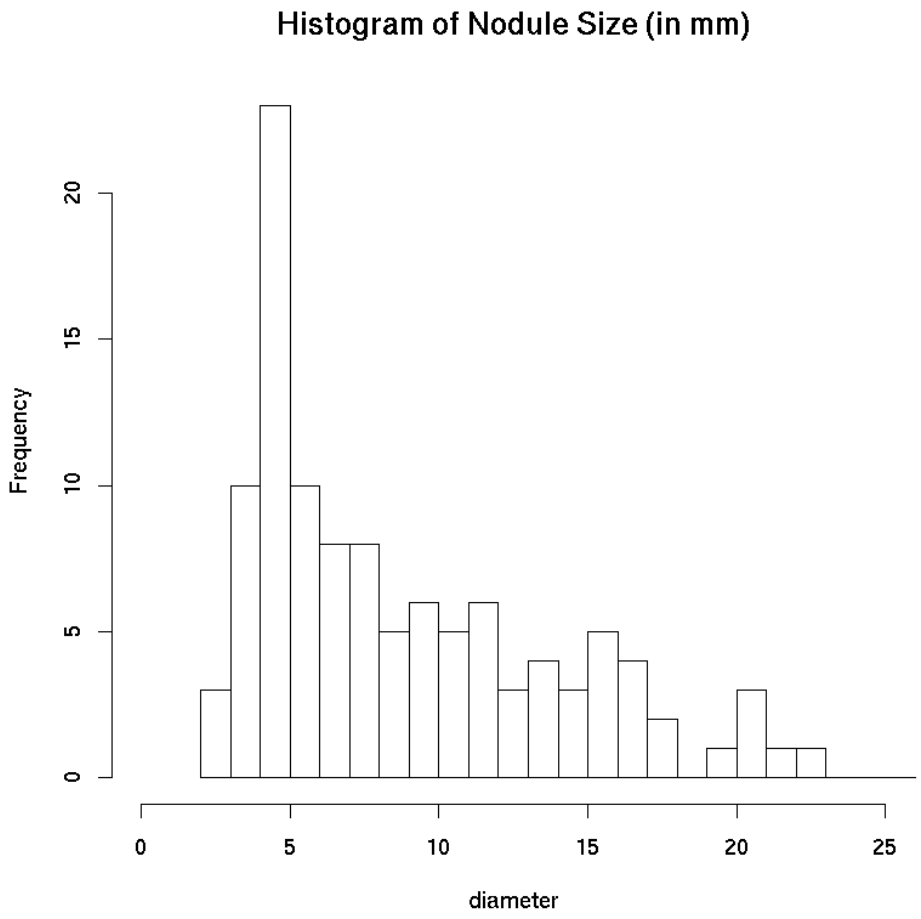


Figure 1. The nodule size distribution of the data set, as computed from the volume estimates of the manual markings and expressed as the diameter of a sphere having the same volume as the estimate.

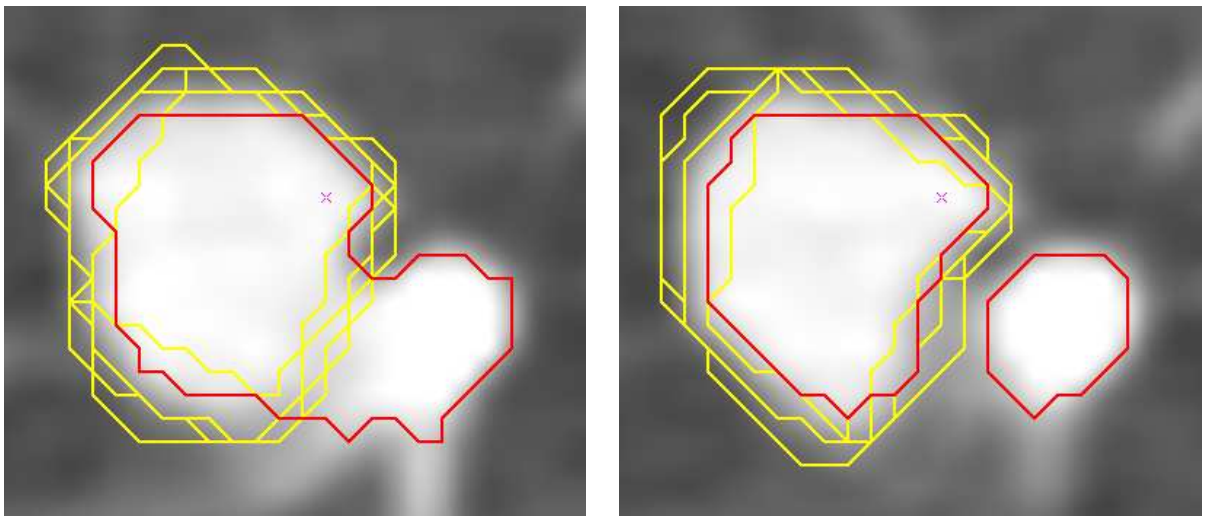
a  b

Figure 2. An example showing two central consecutive slices of a nodule where the automated method, shown in red/dark grey, performed a less than ideal segmentation, failing to remove the vascular structure. The radiologist boundaries, in yellow/light grey, are shown for reference.

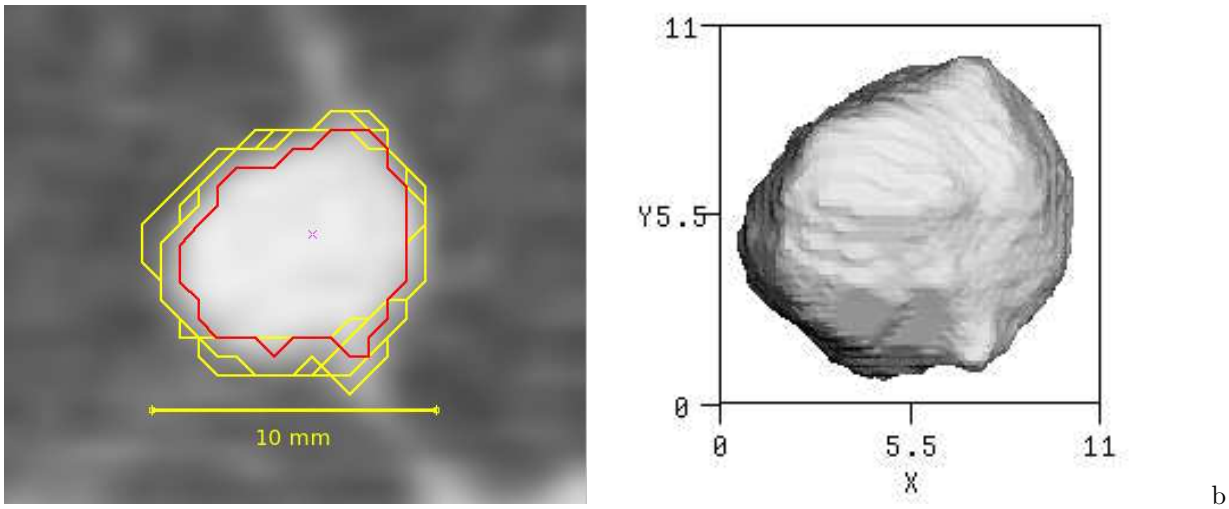
- whether the appearance order of markings in the annotation file is sitewise consistent throughout the file.

All of the above implies that the automated method is the only reader whose values can be tracked among the full set of nodules; for the other readers, only in the presence of multiple nodules in the same scan the estimated values can be attributed to the same readers.

In view of all these points, our first attempt at defining an agreement criterion was to rank the five (absolute) volume estimates and consider the automated method in agreement if its value was not ranked first of fifth. By this citerion, however, when the computer is in disagreement there is always one radiologist who is in disagreement, too. Given the impossibility of defining any kind of reader's profiles (e.g. attributing a constant bias on the performed measures), we preferred looking for a criterion that would give raise of just one disagreement case for nodule. The new criterion ranks the distances from the median estimate and if the automated method is not fifth, then it is considered to be in full agreement with the radiologists. The problem, here, is the set of values used to computed the median because the median could be computed using just the four radiologists' estimates or using the extended set that includes also the computer estimate as the comparison, it could be argued, is actually made among five readers. We decided in favor of the stricter set defined by just the four radiologists' values because the five-value median is biased by the automated method value and we have not yet established that the computer-based values can be considered on a par with the radiologists' ones.

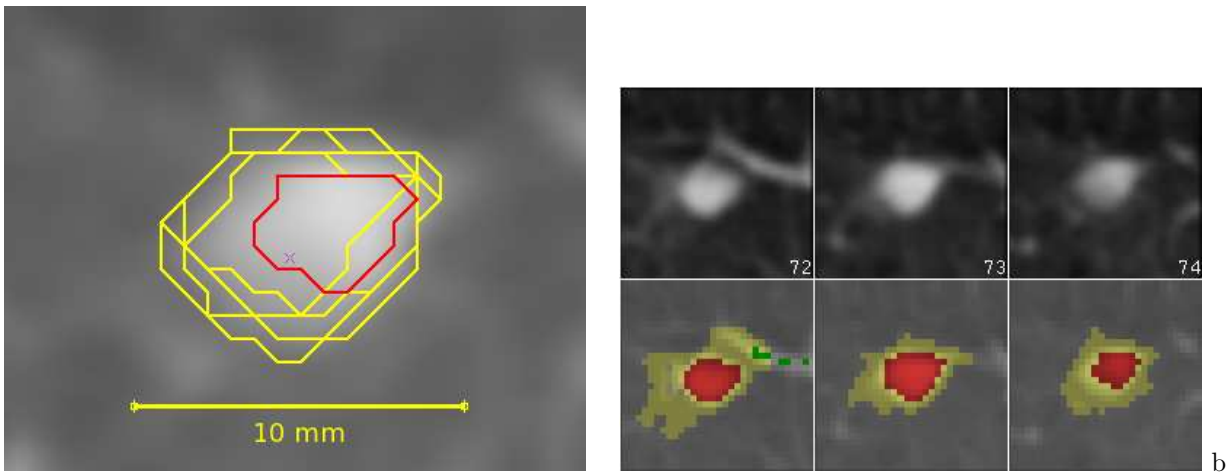
Figure 3 and 4 show examples of the automated method behavior on nodules from the LIDC database. In Figure 3 the segmentation of the standard automated method (red/dark grey boundary) matched the radiologists' performance (yellow/light grey boundaries). Figure 4 shows one case where the augmented automated method better captured the LIDC rules due to the sub-solid component.

The first version of the automated method, segmenting only the solid tissue, resulted in a full agreement in 68.1% of the cases; as expected, this version of automated method underestimates the LIDC criterion volume. The second version of the automated method had a better outcome having an agreement with the radiologists in 72 out of 94 nodules, i.e. 76.6% of the cases. This result can be considered very positive because, even for a human reader, a certain degree of disagreement should be allowed. Given the extent of this study and the intrinsic limitations of the data set, especially as far as the inability of identifying the subsets of markings belonging to each LIDC reader, the 23.4% of the cases where the automated method is at the greatest distance from the reader's median can be considered not a failure, but a reasonable share of the cases where the autonomous reader has expressed its point of view. Consider for a five reader unbiased situation each reader might expect to be at the greatest distance one in 5 times or 20% of the nodules.

a 

b

Figure 3. An example of a solid nodule where the standard automated method (red/dark-grey boundary) was able to match the radiologists' performance (yellow/light-grey boundaries). On the right, a 3D view, along the axial axis, of the whole nodule as segmented by the automated method.

a 

b

Figure 4. An example of a nodule where the sub-solid part prevented the standard automated method from capturing all the valid components for the LIDC specifications. On the upper right, the original frames and, on the lower right, the result of the segmentation by the augmented automated method where the solid part is shown shaded in red/dark grey and the sub-solid part is shown shaded in yellow/light grey.

5. CONCLUSIONS

In this preliminary study, the advances in the automated algorithm for nodule volume measurement proved that computer can effectively assist radiologists in performing the time-consuming task of nodule volumetric measurement. Even if a relatively small number of nodules were not segmented or had major segmentation problems, the deterministic nature of the algorithm can overcome the variability in human readers and open the way for shorter time intervals needed for the analysis of sequential scans. This has important consequences both when the effectiveness of a therapy or the malignancy of a nodule is to be determined in a timely manner. For the future our aim is to improve the automated method on two fronts: on the one hand, raising its robustness with respect to complex or attached nodules and, on the othe hand, making its estimates even closer to the radiologists' ones.

REFERENCES

1. C. R. Meyer, T. D. Johnson, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, B. F. Mullan, D. F. Yankelevitz, E. J. R. van Beek, S. G. Armato III, M. F. McNitt-Gray, A. P. Reeves, D. Gur, C. I. Henschke, E. A. Hoffman, P. H. Bland, G. Laderach, R. Pais, D. Qing, C. Piker, J. Guo, A. Starkey, D. Max, B. Y. Croft, and L. P. Clarke, "Evaluation of lung MDCT nodule annotation across radiologists and methods," *Academic Radiology* **13**, pp. 1254–1265, 2006.
2. A. P. Reeves, A. M. Biancardi, T. V. Apanasovich, C. R. Meyer, H. MacMahon, E. J. van Beek, E. A. Kazerooni, D. Yankelevitz, M. F. McNitt-Gray, G. McLennan, S. G. Armato III, C. I. Henschke, D. R. Aberle, B. Y. Croft, and L. P. Clarke, "The lung image database consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements," *Academic Radiology* **14**, pp. 1475–1485, Dec 2007.
3. J. P. Ko, H. Rusinek, E. L. Jacobs, J. S. Babb, M. Betke, G. McGuinness, and D. P. Naidich, "Small pulmonary nodules: Volume measurement at chest CT – phantom study," *Radiology* **228**, pp. 864–870, September 2003.
4. J.-M. Kuhnigk, V. Dicken, L. Bornemann, D. Wormanns, S. Krass, and H.-O. Peitgen, "Fast automated segmentation and reproducible volumetry of pulmonary metastases in CT-scans for therapy monitoring," in *Lecture Notes in Computer Science*, **3217**, pp. 933–941, Medical Image Computing and Computer-Assisted Intervention, Springer-Verlag GmbH, 2004.
5. K. Okada, D. Comaniciu, and A. Krishnan, "Robust anisotropic gaussian fitting for volumetric characterization of pulmonary nodules in multislice CT," *IEEE Transactions on Medical Imaging* **24**, pp. 409–423, March 2005.
6. L. R. Goodman, M. Gulsun, L. Washington, P. G. Nagy, and K. L. Piacsek, "Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements," *American Journal of Roentgenology* **186**, pp. 989–994, April 2006.
7. M.-P. Revel, A. Merlin, S. Peyrard, R. Triki, S. Couchon, G. Chatellier, and G. Frija, "Software volumetric evaluation of doubling times for differentiating benign versus malignant pulmonary nodules," *American Journal of Roentgenology* **187**, pp. 135–142, July 2006.
8. A. Reeves, A. Chan, D. Yankelevitz, C. Henschke, B. Kressler, and W. Kostis, "On measuring the change in size of pulmonary nodules," *IEEE Transactions on Medical Imaging* **25**, pp. 435–450, April 2006.
9. National Institutes of Health, "Lung image database resource for imaging research." http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-01-001.html, April 2000. Accessed Jul 22nd, 2007.
10. S. G. Armato, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft, L. P. Clarke, and the Lung Image Database Consortium Research Group, "Lung image database consortium: developing a resource for the medical imaging research community," *Radiology* **232**(3), pp. 739–748, 2004.
11. M. F. McNitt-Gray, S. G. Armato III, C. R. Meyer, A. P. Reeves, G. McLennan, R. C. Pais, J. Freymann, M. S. Brown, R. M. Engelmann, P. H. Bland, G. E. Laderach, C. Piker, J. Guo, Z. Towfic, D. P.-Y. Qing, D. F. Yankelevitz, D. R. Aberle, E. J. van Beek, H. MacMahon, E. A. Kazerooni, B. Y. Croft, and L. P. Clarke, "The lung image database consortium (LIDC) data collection process for nodule detection and annotation," *Academic Radiology* **14**, pp. 1464–1474, Dec 2007.

12. R. S. Breiman, J. W. Beck, M. Korobkin, R. Glenny, O. E. Akwari, D. K. Heaston, A. V. Moore, and P. C. Ram, "Volume determinations using computed tomography," *American Journal of Roentgenology* **138**(2), pp. 329–333, 1982.
13. W. J. Kostis, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical ct images," *IEEE Transactions on Medical Imaging* **22**, pp. 1259–1274, Oct. 2003.