Int J CARS
DOI 10.1007/s11548-009-0401-3

ORIGINAL ARTICLE

A comparison of ground truth estimation methods

Alberto M. Biancardi · Artit C. Jirapatnakul · Anthony P. Reeves

Received: 10 January 2009 / Accepted: 17 June 2009
© CARS 2009

Abstract
Purpose Knowledge of the exact shape of a lesion, or ground truth (GT), is necessary for the development of diagnostic tools by means of algorithm validation, measurement metric analysis, accurate size estimation. Four methods that estimate GTs from multiple readers' documentations by considering the spatial location of voxels were compared: thresholded Probability-Map at 0.50 ($TPM_{0.50}$) and at 0.75 ($TPM_{0.75}$), simultaneous truth and performance level estimation (STAPLE) and truth estimate from self distances (TESD).
Methods A subset of the publicly available Lung Image Database Consortium archive was used, selecting pulmonary nodules documented by all four radiologists. The pair-wise similarities between the estimated GTs were analyzed by computing the respective Jaccard coefficients. Then, with respect to the readers' marking volumes, the estimated volumes were ranked and the sign test of the differences between them was performed.
Results (a) the rank variations among the four methods and the volume differences between STAPLE and TESD are not statistically significant, (b) $TPM_{0.50}$ estimates are statistically larger (c) $TPM_{0.75}$ estimates are statistically smaller (d) there is some spatial disagreement in the estimates as the one-sided 90% confidence intervals between $TPM_{0.75}$ and $TPM_{0.50}$, $TPM_{0.75}$ and STAPLE, $TPM_{0.75}$ and TESD, $TPM_{0.50}$ and STAPLE, $TPM_{0.50}$ and TESD, STAPLE and TESD, respectively, show: [0.67, 1.00], [0.67, 1.00], [0.77, 1.00], [0.93, 1.00], [0.85, 1.00], [0.85, 1.00].
Conclusions The method used to estimate the GT is important: the differences highlighted that STAPLE and TESD, notwithstanding a few weaknesses, appear to be equally viable as a GT estimator, while the increased availability of computing power is decreasing the appeal afforded to TPMs. Ultimately, the choice of which GT estimation method, between the two, should be preferred depends on the specific characteristics of the marked data that is used with respect to the two elements that differentiate the method approaches: relative reliabilities of the readers and the reliability of the region boundaries.

Keywords CAD development · Algorithm validation · Volumetric measurement · Diagnosis · Response to therapy

A. M. Biancardi (✉) · A. C. Jirapatnakul · A. P. Reeves
Cornell University, 397 Rhodes Hall, Ithaca, NY, USA
e-mail: amb284@cornell.edu

Introduction

For lung nodules, estimation of growth rates or 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 [14,19] 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, mandating 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 [8,11,12,17,18,20]. The difficulty of this approach is now shifted toward the need to calibrate and to validate such methods

Published online: 09 December 2009

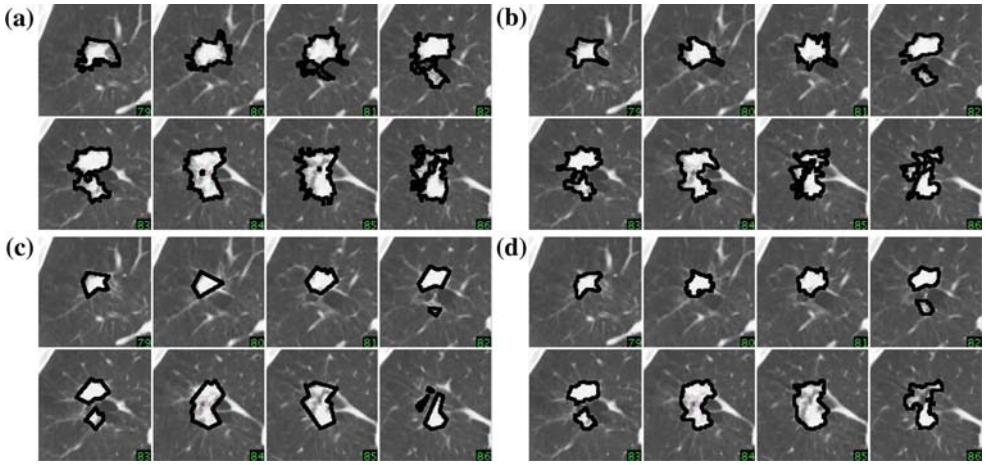

Fig. 1 An example of nodule documentation provided by the LIDC database. In this case, a montage of the central slices of a nodule are displayed and, overlaid, the radiologists' markings of the nodule boundary. Each montage set (**a–c**), and (**d**) shows the markings of one radiologist

and, currently, the only accepted source for the definition of a ground truth (GT) is based on nodule boundary markings performed by expert radiologists.

Scans from the Lung Image Database Consortium [16], one of the answers to this need, were used, given the availability of annotated nodules for the computation of a ground truth (GT) against which automated methods can be tested. In fact, being able to access multiple readers' markings makes it possible to generate an estimate of the actual lesion that (a) takes into account all those different markings and (b) is expected to be a closer representation of the actual lesion region because it aims at minimizing the subjectivity of each reader's marking.

The initial problem of knowing the lesion actual shape, however, has not disappeared: it is now turned into the evaluation of the methods by which the ground truth is estimated as those methods have become a critical part of the validation process. A previous study [3] performed a preliminary comparison between Probability-Maps thresholded at 0.50 and simultaneous truth and performance estimation; the analysis of this paper has been extended to encompass four GT estimation methods: thresholding of probability maps at 0.50 ($TPM_{0.50}$) and 0.75 ($TPM_{0.75}$), simultaneous truth and performance estimation (STAPLE), and truth estimate from self distances (TESD).

Materials

The comparison was performed on whole-lung CT scans provided by the LIDC archive [15]. The LIDC process model [1,13] specifies that each scan is assessed by four experienced thoracic radiologists and that, for nodules 3 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. One of the key tenets of the LIDC process model is the absence of an explicit consensus stage where only one region is provided as the GT, independently from the actual number of radiologists that supplied the initial boundaries. Instead a double reading process, performed by every radiologist, was established and up to four boundaries are provided for each documented nodule corresponding to the radiologists' individual markings. Only nodules marked by all four LIDC radiologists were selected from the LIDC database. Figure 1 shows a montage of the central slices of a nodule and, overlaid, its documentation provided by each of the four radiologists in the form of a boundary marking (a–d).

For this paper 85 whole-lung CT scans were available. All of the scans were acquired from multi-detector row CT scanners with pixel size ranging from 0.508 to 0.762 mm (average 0.64 mm) and an axial slice thickness ranging from 0.75 to 3.00 mm (average 2.07 mm, median 1.80 mm). The tube current ranged from 40 to 422 mA (average 134.4 mA, median 75 mA), tube voltage range was for more than half of the cases 120 kVp with the remaining ones having voltages between 130 and 140 kVp. A total of 35 nodules documented by all four readers were selected. The median sizes, expressed as volumetric-based diameters derived from the manual markings, ranged from 4.4 to 23.8 mm (mean 11.1 mm, median 7.46 mm).

An important aspect is that the LIDC data are anonymized; the anonymization is performed not only on the DICOM image data, where every tag that may lead to the identification of the original subject is transformed or removed, but also on all the XML documentation that is attached to each scan. Therefore, a number of key aspects cannot be known or taken for granted, such as:

- which sites produced the markings (there are five sites cooperating to the LIDC, but only four readers [13]);
- whether a site has multiple readers and, therefore, what is the actual number of radiologists that contributed the annotations.

Since for each scan all the nodule documentations are grouped by reader, the only known element is that the actual readers can be tracked across that limited subset when multiple nodules are present in one scan. However, this information was not used in order to avoid any disparities in the evaluation of nodules.

Ground truth estimators

There are several approaches for the estimation of ground truth from several expert readers' markings. In this study, four methods that perform this estimate by considering the spatial location of voxels were considered: two based on thresholded Probability-Maps (TPM), simultaneous truth and performance level estimation (STAPLE), and truth estimate from self distances (TESD).

Thresholded probability maps

In a Probability-Map [14], the value of a voxel is the weighted average of the values of the voxel in each reader's segmentations. For example, if a voxel is labeled as being part of the lesion by three out of the four readers, the voxel will have a value of 0.75. Using this method, voxels present in all of the readers' segmentations will have a value of 1.0, voxels present in none of the segmentations have a value of 0, and voxels in some, but not all, of the segmentations will have a value of 0.25, 0.50, or 0.75. To generate an estimated GT, the Probability-Map may be thresholded at a particular value. In this study, the Probability-Maps are thresholded at 0.50 (i.e. 50%) to give a thresholded probability-map ($TPM_{0.50}$) that represent the regions marked by two or more readers, and at 0.75 (i.e. 75%) to give a thresholded probability-map ($TPM_{0.75}$) that represent the regions marked by three or more readers.

Simultaneous truth and performance level estimation

The second approach considered in this study is an algorithm proposed by Warfield et al. [23], simultaneous truth and performance level estimation (STAPLE). In this method, the true GT is treated as a hidden variable and therefore not directly observable. Since the true GT is unknown, reader performance is also unknown. The first stage of this method estimates both the GT and reader performance simultaneously using an expectation–maximization (EM) algorithm. These intermediate results are an image similar to a Probability-Map, with the value of each voxel representing the probability for that voxel to be part of the GT, and the estimated sensitivity and specificity for each reader. In this study, our implementation of the STAPLE algorithm is loosely based on the version from the National Library of Medicine Insight Segmentation and Registration Toolkit (ITK) [9] to which we added the graph cut post-processing. The STAPLE algorithm can be initialized by assuming sensitivity and specificity values for each reader, or by assuming an initial GT. Following the authors' indications [23], all readers are given the same initial sensitivity and specificity as their true quality is unknown. Thus, the initial ground truth estimate is an equally weighted average of all of the reader segmentations. Another parameter of the algorithm is the selection of a function for the prior probability that a pixel is included in the GT segmentation. A reasonable value to use is the relative proportion of segmented pixels in the true segmentations [23] and this value is used in this study. The final step of the STAPLE method is based on the construction of an hidden Markov random field and its use to generate the actual final estimate. In this step the voxel independence assumption is removed and the respective relationships of each voxel with its neighbors are used to regularize the EM estimate. The finding of the optimal solution, following [23], was realized as a max-flow min-cut by tailoring the implementation provided by Boykov and Kolmogorov [4] of the algorithm formulated by Ford and Fulkerson [7].

Truth estimate from self distances

The third approach, TESD, was proposed by Biancardi and Reeves [2]. In this algorithm, each marking is processed to produce a three-dimensional binary occupancy region where voxels are given a value of 1 to mean that, according to the reader evaluation, that voxel is part of the lesion; zero-valued voxels are considered outside the lesion. The set of occupancy regions, derived from the readers' markings, $\{R_i; 1 \leq i \leq 4\}$ is then analyzed as follows:

- Every region R_i is processed to compute the exact signed three-dimensional euclidean distance transform D_i in millimeter. Inside voxels have positive distances that increase when moving from the border toward the lesion center (assuming no holes are present) while outside voxels have negative distances that decrease (increase in absolute value) when moving further away from the border;
- Distance values D_i are used to create the labeled voting maps L_i where each voxel is labeled, as shown in Fig 2, into four categories according to its signed distance-transform (implemented as an exact distance in millimeter with a linear complexity as described in [6]): inner core, inside border, outside border, and outer space;

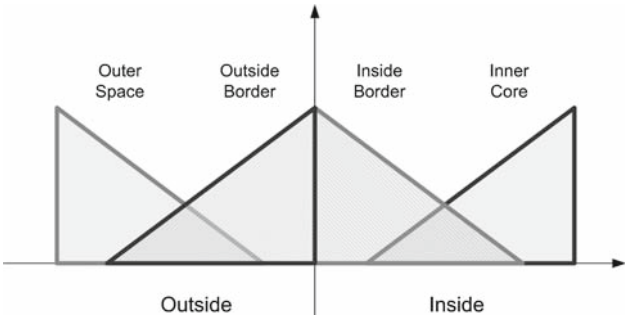

Fig. 2 A graph showing the mapping from distance transform to weighted categories in TESD

- The labeled maps L_i are then combined by applying, voxel by voxel, a center of gravity method to produce a defuzzyfied map with crisp values that is then thresholded, at an empirically determined value, to provide the final binary estimate.

The motivations behind the development of this GT estimation method were explained in the need of capturing the uncertainty of the reader in determining the exact location of the borders and of giving equal weights to experienced readers who may disagree on the exact placement of the lesion border, but that largely agree on the location of the lesion main core. Accordingly, two primary requirements were elicited:

- the importance of not giving more reliability to one or more readers to the disadvantage of others, and
- the ability of performing the estimate by analyzing the neighborhood of each voxel and the co-location of all the marked regions, evaluating also how close they are and not only how much they overlap.

The first requirement constrained the development of the method by considering as not adequate both those label weighting strategies that would depend on the shape of the marked region and those defuzzyfication methods that would treat their input arguments asymmetrically. The second requirement led to a key aspect of TESD: while other methods give non-zero values only to voxels inside each reader's marking, in TESD also outside voxels have values different from zero in order to capture both the shape of the lesion and the closeness of one reader's marked voxels to the other reader's ones.

It is worth underlining how the GT estimation by TESD is not a shape averaging [21] because, even if the first step is similar with the creation of signed distance-transform maps, TESD does not use the distance values directly, but transforms them into one of four categories.

Methods

After determining the GT estimates for all of the four methods, volumes for the estimates were computed by counting the number of nodule pixels in each of the image slices and then multiplying their sum by the voxel volume [5]—a method frequently used in CAD/CADx tools. When expressing the volumes in the uni-dimensional scale space, like the one used by RECIST [22], the diameter d of the equivalent sphere was used, i.e. the diameter of a sphere having the same volume as the estimate: $d = 2\sqrt[3]{\frac{3v}{4\pi}}$.

The first point of the comparison was carried out by computing the rank of the estimated volumes against the four volumes of the marked regions, i.e. creating a list with the five volume values, sorting them in increasing order and record the position of the volume estimates in the sorted list. It is expected that the estimates never take the first (smallest) or the last (largest) positions; if we define a value to be in agreement with the others when it is in the inner position, this means that we expect the estimate to never disagree with respect to the readers' markings. Then, for each pair of estimators, two indicators of the pair-wise similarities were computed: the Fisher sign test of the differences between the respective estimated volumes and the one-sided 90% confidence interval (CI) of the Jaccard coefficients between the respective estimated regions. The Fisher sign test is used to verify that, given a set of measurement pairs $\{x_i, y_i\}$, x_i and y_i are equally likely to be larger than the other by testing the null hypothesis for the sign of the differences to follow the binomial distribution with probability $p = 0.5$. Given two sets X and Y, the Jaccard coefficient [10] measures the amount of overlap between the two sets and is defined as:

$$J = \frac{|X \cap Y|}{|X \cup Y|}$$

The analysis was also extended to include an evaluation of the differences between the TPM and the final STAPLE regions with STAPLE intermediate results, computed after the EM stage, by generating a binary GT estimate by thresholding the stage output map at a probability of one half (50%).

It is important to underline that all the GT estimations were performed considering each nodule separately without extending the reader performance level evaluation to more than a single nodule because no reader can be tracked among the full set of nodules due to the anonymization of the LIDC data as explained above. The possibility to track readers among the multiple nodules of one case was rejected because it would have created a disparity in the evaluation of nodules.

Results

Except for $TPM_{0.75}$, none of the other methods had any nodules in the rank extremes as expected. The distributions for the other three ranks, from the smallest to the largest, for $TPM_{0.50}$ was 1,18,16, for STAPLE was 6, 20, 9, and for TESD was 4, 22, 9. For $TPM_{0.75}$, the distribution of across all the five ranks was 4, 27, 4, 0, 0.

The spatial agreement between the GT estimates of the $TPM_{0.50}$ and the $TPM_{0.75}$ had a one-sided 90% CI of [0.67, 1.00]: this was reflected into relative volume differences ranging from 8.88 to 107.1% (average 27.5%, median 22.2%) or, expressing them in a uni-dimensional scale space, into a range from 2.88 to 34.1% (average 8.9%, median 6.9%). The sign test on the difference between the volume estimations had a p value <0.001.

The spatial agreement between the GT estimates of the $TPM_{0.75}$ and STAPLE had a one-sided 90% CI of [0.67, 1.00]: this was reflected into relative volume differences ranging from 2.41 to 107.1% (average 24.4%, median 19.6%) or, expressing them in a uni-dimensional scale space, into a range from 0.8 to 33.9% (average 8.0%, median 6.2%). The sign test on the difference between the volume estimations had a p value <0.001.

The spatial agreement between the GT estimates of the $TPM_{0.75}$ and TESD had a one-sided 90% CI of [0.77, 1.00]: this was reflected into relative volume differences ranging from 6.6 to 76.8% (average 20.5%, median 17.2%) or, expressing them in a uni-dimensional scale space, into a range from 2.1 to 20.9% (average 6.3%, median 5.6%. The sign test on the difference between the volume estimations had a p value <0.001.

The spatial agreement between the GT estimates of the $TPM_{0.50}$ and STAPLE had a one-sided 90% CI of [0.93, 1.00]: this was reflected into relative volume differences ranging from −0.3 to 12.0% (average 2.6%, median 1.3%) or, expressing them in a uni-dimensional scale space, into a range from −0.1 to 4.2% (average 0.9%, median 0.4%). The sign test on the difference between the volume estimations had a p value < 0.001.

The spatial agreement between the GT estimates of the $TPM_{0.50}$ and TESD had a one-sided 90% CI of [0.85, 1.00]: this was reflected into relative volume differences ranging from −2.0 to 17.0% (average 5.1%, median 3.2%) or, expressing them in a uni-dimensional scale space, into a range from −0.7 to 6.0% (average 1.8%, median 1.1%). The sign test on the difference between the volume estimations had a p value < 0.001.

Table 1 A summary of the one-sided 90% CI of the Jaccard coefficients

	$TPM_{0.75}$	$TPM_{0.50}$	STAPLE
$TPM_{0.50}$	[0.67, 1.00]		
STAPLE	[0.67, 1.00]	[0.93, 1.00]	
TESD	[0.77, 1.00]	[0.85, 1.00]	[0.85, 1.00]

The spatial agreement between the GT estimates of the STAPLE and TESD had a one-sided 90% CI of [0.85, 1.00]: this was reflected into relative volume differences ranging from −15.9 to 17.2% (average 2.4%, median 1.3%) or, expressing them in a uni-dimensional scale space, into a range from −5.0 to 6.1% (average 0.9%, median 0.4%). The sign test on the difference between the volume estimations had a p value of 0.19 and therefore the size differences of the estimates are not statistically significant.

A summary of these results is presented in Tables 1 and 2 A first example, Fig. 3, shows in detail the marked regions (r1) to (r4), the intermediate stages for the methods, (i1) to (i3), and the final estimates, (e1) to (e4). Another example of the GT estimates, spanning multiple slices, is shown in Fig. 4 as a set of tiled images, one for each of the consecutive axial slices that make up the marked and estimated regions. In the left column, each row displays the region marked by one of the readers; in the right column, each row displays the estimation results for each one of the methods: (e) $TPM_{0.75}$, (f) $TPM_{0.50}$, (g) STAPLE, (h) TESD.

Discussion

The evaluation of the rank distributions showed that the $TPM_{0.75}$ estimator is clearly providing estimates that are too small when compared to the reader's volumes, while the estimates for all of the other methods were bracketed by the values of the readers' marked regions. This was expected because, even if the estimate is based on the spatial evaluation of the regions, the methods aim at finding a consensus portion that will probably be larger than the smallest marked region and smaller than the largest marked region. Even so, the distribution of the $TPM_{0.50}$ estimator appears to

Table 2 A summary of the relative differences in the volume estimates of the methods: minimum, maximum and, in parethesis, average and median

	$TPM_{0.75}$	$TPM_{0.50}$	STAPLE
$TPM_{0.50}$	8.88 to 107.1% (27.5, 22.2%)		
STAPLE	2.41 to 107.1% (24.4, 19.6%)	−0.3 to 12.0% (2.6, 1.3%)	
TESD	6.6 to 76.8% (20.5, 17.2%)	−2.0 to 17.0% (5.1, 3.2%)	−15.9 to 17.2% (2.4, 1.3%)

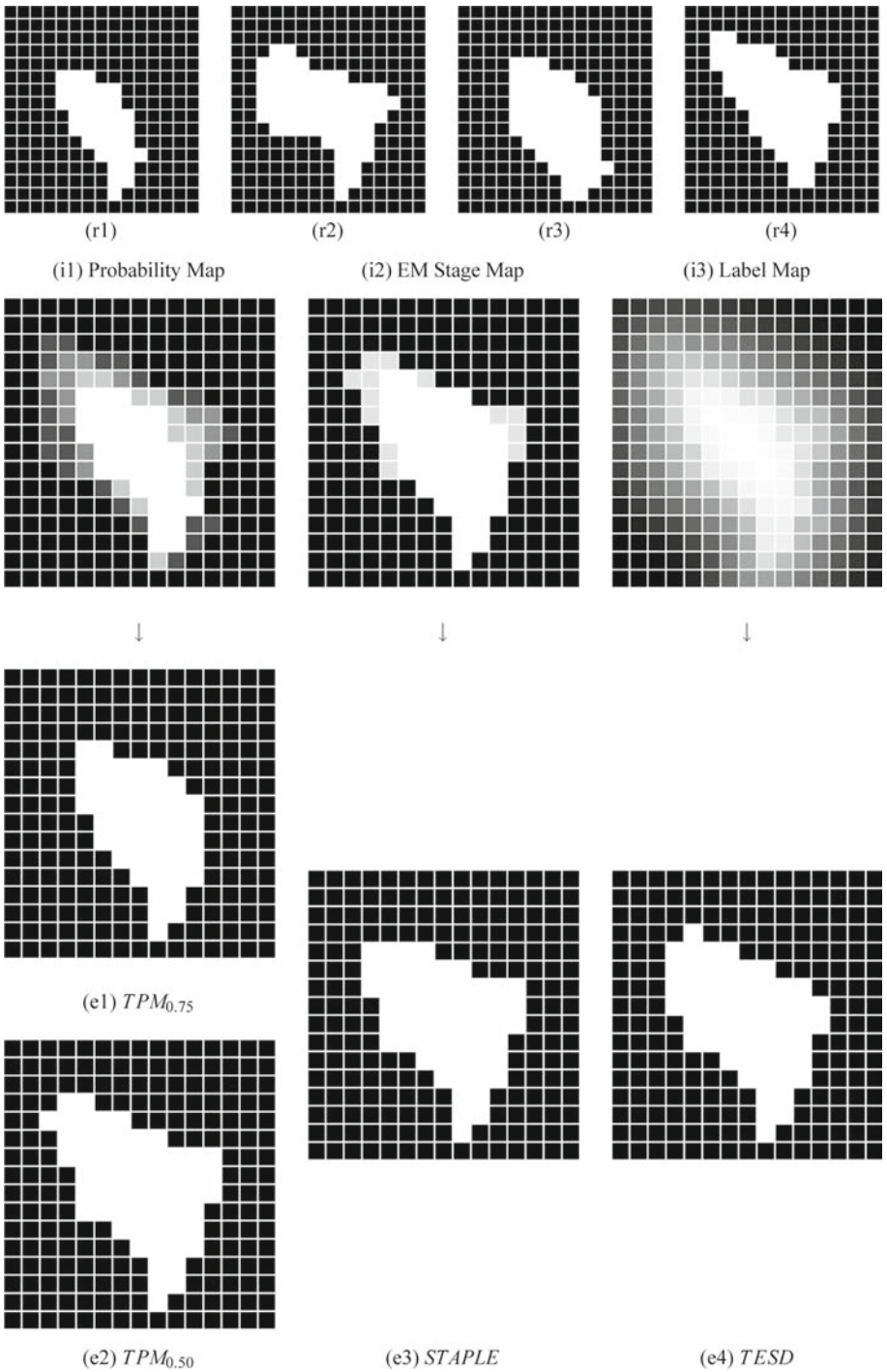

Fig. 3 An estimation example of a single slice extracted from the full set of markings and results. Readers' marked regions are displayed as (*r1*)–(*r4*), normalized intermediate results for the three methods as (*i1*)–(*i3*), where white corresponds to the highest value, and final binary estimates as (*e1*)–(*e4*)

be skewed toward the larger side of the size spectrum, having almost as many estimates in the larger fourth rank as it has in the middle third rank.

The Jaccard coefficients showed a certain degree of disagreement between STAPLE and $TPM_{0.50}$ and the sign tests showed that STAPLE volumes are almost always lesser than $TPM_{0.50}$ ones, indicating a systematic difference between the two methods and this motivated us to extend our analysis also to the output of STAPLE EM stage. The EM estimate is included in the $TPM_{0.50}$ estimate (all the voxels marked by at least two readers) and includes the $TPM_{0.75}$ estimate (all the voxels marked by at least three readers): hence only the voxels marked by exactly two readers are responsible for making the difference between $TPM_{0.50}$ and the EM estimate. The analysis showed that the effect of the max-flow min-cut optimization, that brings to the final results for STAPLE, is

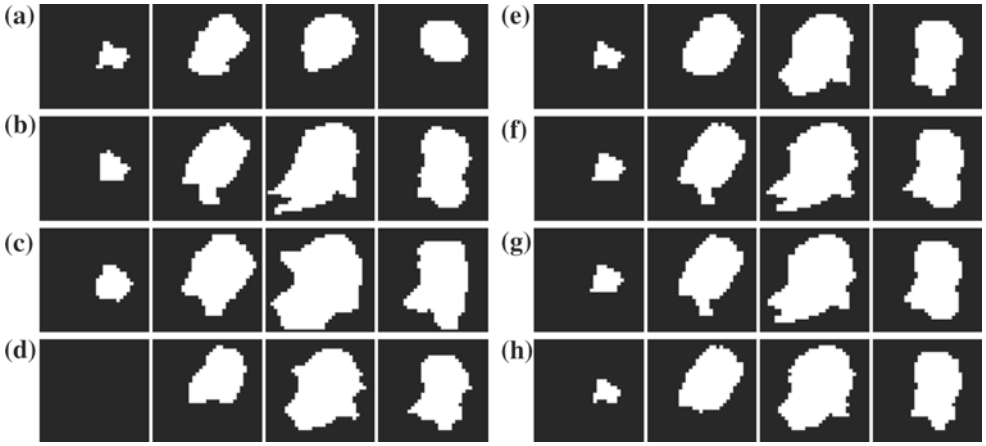

Fig. 4 An example of GT estimations. On the *left column tiles* (**a**–**d**) show the readers' markings; on the *right column* the GT estimates of (**e**) $TPM_{0.75}$, (**f**) $TPM_{0.50}$, (**g**) STAPLE, and (**h**) TESD are shown

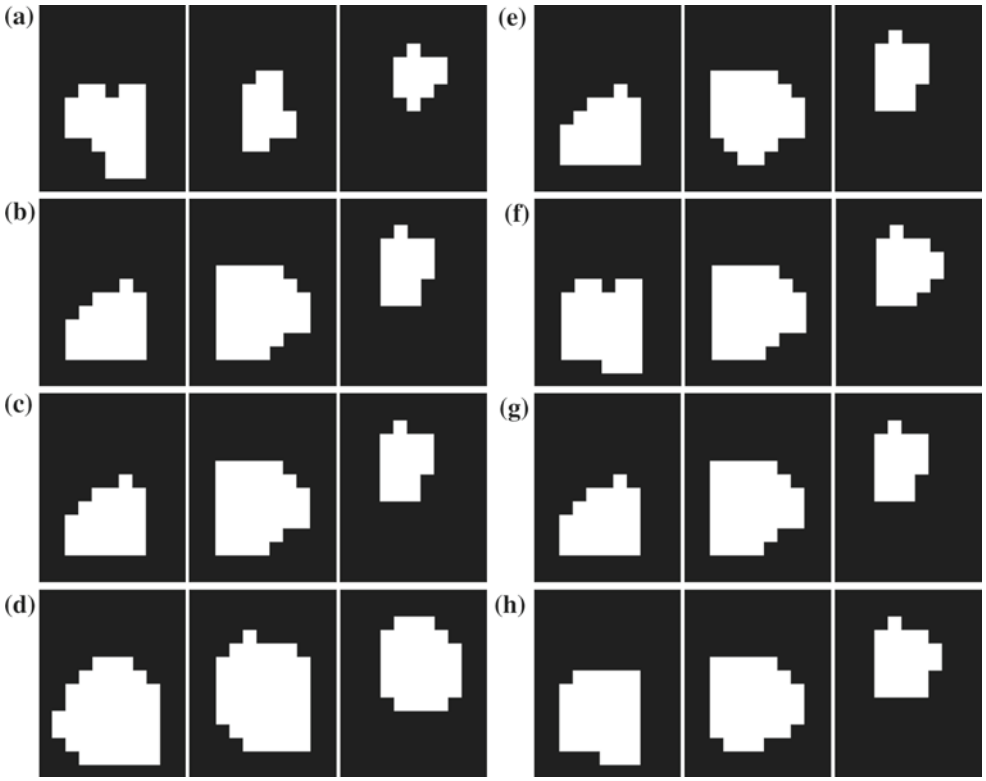

Fig. 5 An example of nodule where two overlapping boundaries are present. On the *left column tiles* (**a**) to (**d**) show the readers' markings; on the *right column* the GT estimates of (**e**) $TPM_{0.75}$, (**f**) $TPM_{0.50}$, (**g**) STAPLE, and (**h**) TESD are shown

that only voxels marked by exactly two readers are further removed, while very few voxels marked by just one reader may get added.

In only 6 out of 35 nodules (17%), the EM estimate was different from the $TPM_{0.50}$ estimate and that prompted further investigation. A key assumption in the theoretical foundation of both methods is voxel spatial independence: the first part of the STAPLE estimation is performed by the EM stage and then, as we said previously, the post-processing by a hidden Markov random field takes into account spatial dependence. One of the other differences between the TPM methods and STAPLE concerns each reader' reliability: the $TPM_{0.50}$ and $TPM_{0.75}$ assume each reader as equally reliable and blindly selects all the voxels marked by at least two readers, whereas STAPLE weights each reader's reliability according to her agreement with the others. However, while

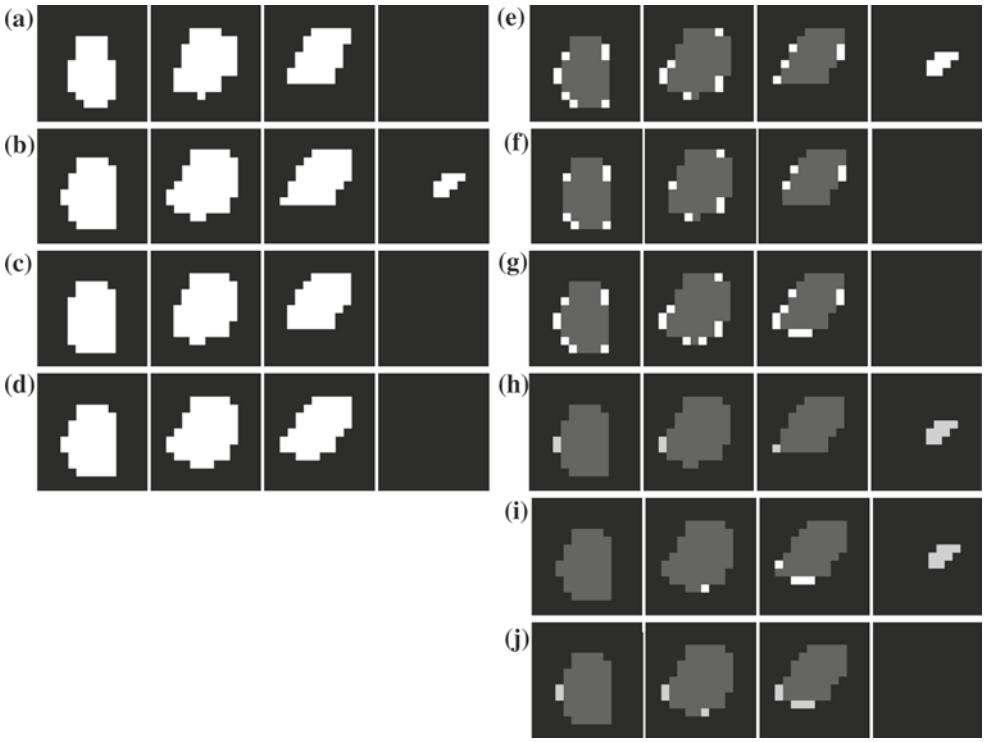

Fig. 6 An example of nodule where STAPLE and TESD estimates mostly agree and TPM estimates differ from the other two methods. In the *left column*, the GT estimates of (**a**) $TPM_{0.75}$, (**b**) $TPM_{0.50}$, (**c**) STAPLE, and (**d**) TESD are shown. The *right column* displays the pairwise differences between those regions: (**e**) $TPM_{0.75}$ and $TPM_{0.50}$, (**f**) $TPM_{0.75}$ and STAPLE, (**g**) $TPM_{0.75}$ and TESD, (**h**) $TPM_{0.50}$ and STAPLE, (**i**) $TPM_{0.50}$ and TESD, (**j**) TESD and STAPLE. Common regions are shown in *dark gray*, voxels belonging to the first estimate only are in *white*, and voxels voxels belonging to the second estimate only are in *light gray*

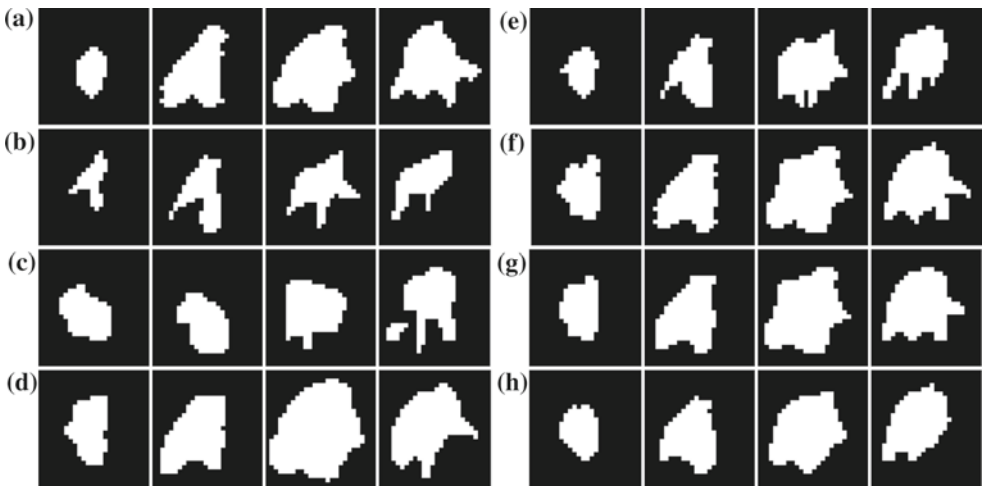

Fig. 7 An example of nodule where $TPM_{0.75}$ and TESD estimates differs from the other two methods. On the *left column* tiles (**a**–**d**) show the readers' markings; on the *right column*, the GT estimates of (**e**) $TPM_{0.75}$, (**f**) $TPM_{0.50}$, (**g**) STAPLE, and (**h**) TESD are shown

this feature tries to capture readers' agreements and disagreements, it may also be affecting the results. In those six case, brought to our attention by the EM result analysis, STAPLE shows the following behavior: (a) when there are two readers' markings almost coincidental, which happened in 4 out of 35 cases (i.e. more than 11% of the available documented nodules), the estimate includes the voxels belonging to these two and excludes almost every other one, even if marked by both the other two readers, as shown in Fig. 3; (b) in the other two cases the reader that drew the largest region is assigned a higher sensitivity (because the summation of non-zero contributions spans a larger number of voxels), which

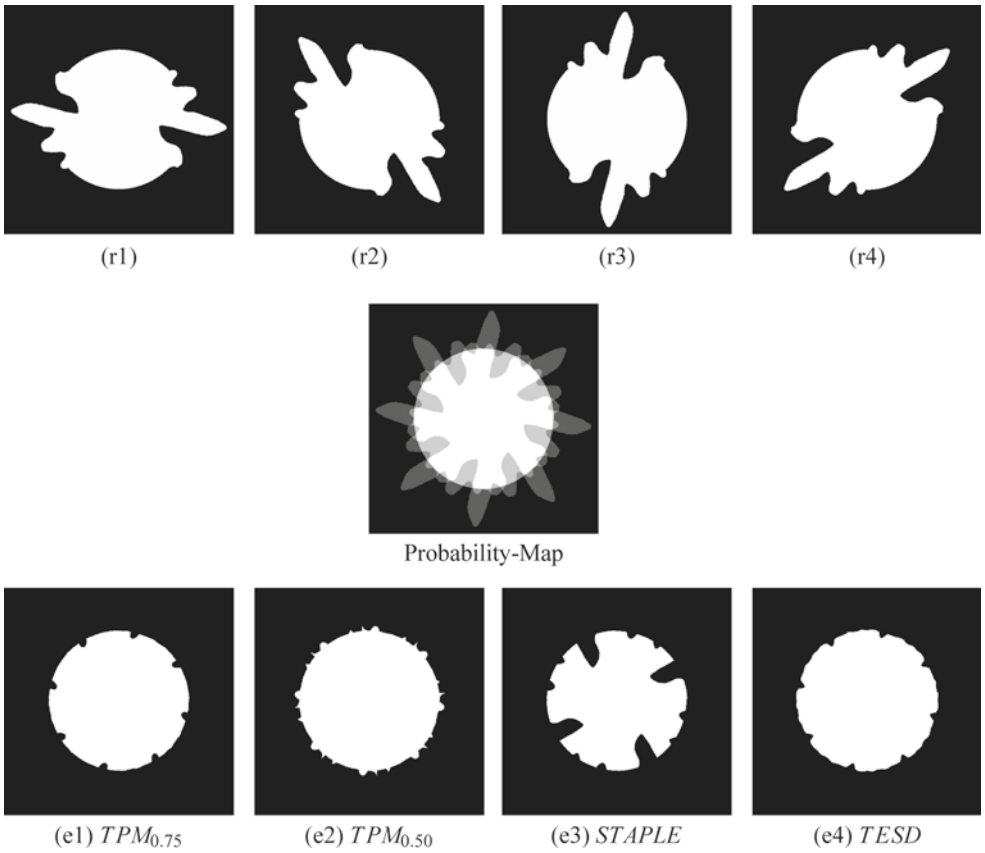

Fig. 8 The ad hoc example to highlight the underlying differences between the methods based on equally reliable readers and exact borders *(e1)* $TPM_{0.75}$ and *(e2)* $TPM_{0.50}$, variable reliability and exact borders *(e3)* STAPLE, and equally reliable readers with fuzzy border definition *(e4)* TESD. The *top row* shows the simulated reader's markings, the *middle row* the probability map and the *lowest row* the estimates

in turn increases the probability of belonging to the GT to all the voxels marked by that reader and excludes the other pairs that are in disagreement. In one case, both factors were presents, almost coinciding marks, but not identical, and all of the voxels that were part of the largest of the two were included in the estimated GT.

Figure 5 shows an example when there are two overlapping boundaries (b) and (c): in this case STAPLE estimate (f) is determined by that region, while the other two estimates, (e) $TPM_{0.50}$ and (g) TESD, include voxels marked by the other readers.

The Jaccard coefficient between TESD and the other two methods showed, as expected, a certain disagreement, while the Fisher sign test showed that STAPLE and TESD results are close, as it has to be, taking both somehow into accounts voxels spatial dependence. The difference between $TPM_{0.50}$ and TESD is clear since, as explained before, $TPM_{0.50}$ is simply the region of the voxels marked by at least two readers, hence, it assumes voxel spatial independence and it does not take in account any other spatial element (shape, voxel position, …). The difference between STAPLE and TESD is due to the fact that STAPLE starts with an intermediate probability map (the result of the EM stage) and then applies what happens to be a geometric regularization, while TESD, by its very definition, produces an estimate that is based on the respective locations of the marked regions and where each voxel outcome is determined by the effects of a limited number of neighboring voxels.

The tendency of estimate larger regions by $TPM_{0.50}$ is displayed in Fig. 6 where, in addition to the three GT estimates, (a), (b) and (c), the pairwise differences between those regions show how the whole portion in the rightmost tile is only included by $TPM_{0.50}$, while the differences between STAPLE and TESD are limited to a few voxels. The estimate by TESD for the nodule in Fig. 7 differs, from the other two methods, especially in the lower right corner of the rightmost tile. In this case, the two factors of co-location and locality of TESD evaluation were responsible for the visible reduction in the result region; slightly less noticeable, but still present, is STAPLE regularization that can be observed when comparing its estimate with $TPM_{0.50}$.

A study of a set of spherical phantom nodules, where known GTs are available, was performed; with the exception of $TPM_{0.75}$, the differences between the methods were

negligible because of the high agreement among readers due to the spherical form of the phantoms, confirming a certain degree of reliability of the methods. Therefore, in order to better highlight the different behaviors of the estimators we developed a set of ad hoc test cases, an example of which is discussed here and is shown in Fig. 8. Similarly to the selection criterion for the LIDC dataset that was analyzed, there are four simulated readers and each of their markings is a rotated version of the same prototype: a region obtained by arbitrary fluctuating its boundary around a circle of a diameter of 311 pixels and an area of 75964.5 pixels^2. The four input markings are shown as images (r1) to (r4) in Fig. 8; their Probability-Map, also displayed in Fig. 8, clearly shows the circle around which the boundaries are drawn. The four results highlight the underlying differences among the methods: $TPM_{0.75}$, (e1), and $TPM_{0.50}$, (e2), are based on the evaluation of equally reliable readers whose marked regions have totally reliable borders; STAPLE, (e3), is based on readers with variable reliability and whose marked regions have totally reliable borders; TESD, (e4), is based on equally reliable readers whose marked regions have fuzzy-defined borders. The number of pixels belonging to the estimated GTs are: for $TPM_{0.75}$ 74457, for $TPM_{0.50}$ 78391, for STAPLE 66162, and for TESD 75905.

Conclusions

A comparison of the GT regions created by $TPM_{0.50}$, STAPLE, and TESD was performed. Major findings of this analysis were: (a) the variations in rank distributions, when comparing the estimates with the marked regions, are not statistically significant, (b) the greater size of $TPM_{0.50}$ estimates with respect to the estimates of the other three methods is statistically significant, (c) the smaller size of $TPM_{0.75}$ estimates with respect to the estimates of the other three methods is statistically significant, (d) the differences in the volume estimates between STAPLE and TESD estimates are not statistically significant, (e) there is spatial disagreement between the estimates of all the method pairs, (e) STAPLE results can be affected either by the closeness between two mark-ups or by larger regions being favored over smaller ones, and (f) TESD provides reasonable estimates, slightly different than the other two methods since it relies on the co-location of radiologists' marked regions and on the nodule in its three-dimensionality. Being the method used to estimate the GT so important, the differences highlighted that STAPLE and TESD, notwithstanding a few weaknesses, appear to be equally viable as a GT estimator, while the increased availability of computing power is decreasing the appeal afforded to $TPM_{0.50}$. Ultimately, while TPMs showed to be not so good, the choice of which GT estimation method between the two more relevant ones (STAPLE and TESD) should be preferred depends on the specific characteristics of the marked data that is used with respect to the two elements that differentiate the approach of the two methods: relative reliabilities of the readers and the reliability of the region boundaries.

References

1. Armato SG, McLennan G, McNitt-Gray MF, Meyer CR, Yankelevitz D, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Reeves AP, Croft BY, Clarke LP (2004) The Lung Image Database Consortium Research Group. Lung image database consortium: developing a resource for the medical imaging research community. Radiology 232(3):739–748
2. Biancardi AM, Reeves AP (2009) TESD: a novel ground truth estimation method. In: SPIE international symposium on medical imaging, vol 7260, pp 72603V–1–8
3. Biancardi AM, Jirapatnakul AC, Fotin S, Apanasovich TV, Reeves AP (2009) An analysis of two ground truth estimation methods. In: SPIE international symposium on medical imaging, vol 7260, pp 72600E–1–8
4. Boykov Y, Kolmogorov V (2004) An experimental comparison of Min-Cut/Max-Flow algorithms for energy minimization in vision. IEEE Trans Pattern Anal Mach Intell 26(9):1124–1137
5. Breiman RS, Beck JW, Korobkin M, Glenny R, Akwari OE, Heaston DK, Moore AV, Ram PC (1982) Volume determinations using computed tomography. Am J Roentgenol 138(2):329–333
6. Felzenszwalb P, Huttenlocher D (2003) Distance transforms of sampled functions. Technical report, Cornell University
7. Ford LR Jr, Fulkerson DR (1956) Maximal flow through a network. Can J Math 8(3):399–404
8. Goodman LR, Gulsun M, Washington L, Nagy PG, Piacsek KL (2006) Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements. Am J Roentgenol 186:989–994
9. Ibanez L, Schroeder W, Ng L, Cates J (2005) The ITK Software Guide, 2nd edn. Kitware. ISBN 1-930934-15-7. http://www.itk.org/ItkSoftwareGuide.pdf
10. Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaudoise Sci Nat 44:223–270
11. Ko JP, Rusinek H, Jacobs EL, Babb JS, Betke M, McGuinness G, Naidich DP (2003) Small pulmonary nodules: volume measurement at chest CT—phantom study. Radiology 228(3):864–870
12. Kuhnigk J-M, Dicken V, Bornemann L, Wormanns D, Krass S, Peitgen HO (2004) Fast automated segmentation and reproducible volumetry of pulmonary metastases in CT-scans for therapy monitoring. In: Lecture notes in computer science, vol 3217, pp 933–941. Medical Image Computing and Computer-Assisted Intervention. Springer GmbH
13. McNitt-Gray MF, Armato SG III, Meyer CR, Reeves AP, McLennan G, Pais RC, Freymann J, Brown MS, Engelmann RM, Bland PH, Laderach GE, Piker C, Guo J, Towfic Z, Qing DP-Y, Yankelevitz DF, Aberle DR, van Beek EJR, MacMahon H, Kazerooni EA, Croft BY, Clarke LP (2007) The lung image database consortium (LIDC) data collection process for nodule detection and annotation. Acad Radiol 14(12):1464–1474
14. Meyer CR, Johnson TD, McLennan G, Aberle DR, Kazerooni EA, MacMahon H, Mullan BF, Yankelevitz DF, van Beek EJR, Armato SG III, McNitt-Gray MF, Reeves AP, Gur D, Henschke CI, Hoffman EA, Bland PH, Laderach G, Pais R, Qing D, Piker C, Guo J, Starkey A, Max D, Croft BY, Clarke LP (2006) Evaluation of lung MDCT nodule annotation across radiologists and methods. Acad Radiol 13:1254–1265

15. National Cancer Institute (2009) National cancer imaging archive. https://imaging.nci.nih.gov/ncia/. Accessed 9 Jan 2009
16. National Institutes of Health (2009) Lung image database resource for imaging research. http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-01-001.html, 2000. Accessed 9 Jan 2009
17. Okada K, Comaniciu D, Krishnan A (2005) Robust anisotropic gaussian fitting for volumetric characterization of pulmonary nodules in multislice CT. IEEE Trans Med Imaging 24(3):409–423
18. Reeves A, Chan A, Yankelevitz D, Henschke C, Kressler B, Kostis W (2006) On measuring the change in size of pulmonary nodules. IEEE Trans Med Imaging 25(4):435–450
19. Reeves AP, Biancardi AM, Apanasovich TV, Meyer CR, MacMahon H, van Beek EJR, Kazerooni EA, Yankelevitz D, McNitt-Gray MF, McLennan G, Armato SG III, Henschke CI, Aberle DR, Croft BY, Clarke LP (2007) The lung image database consortium (LIDC): a comparison of different size metrics for pulmonary nodule measurements. Acad Radiol 14(12):1475–1485
20. Revel M-P, Merlin A, Peyrard S, Triki R, Couchon S, Chatellier G, Frija G (2006) Software volumetric evaluation of doubling times for differentiating benign versus malignant pulmonary nodules. Am J Roentgenol 187:135–142
21. Rohlfing T, Maurer CR (2007) Shape-based averaging. IEEE Trans Image Process 16(1):153–161
22. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG (2000) New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 92(3):205–216
23. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23(7):903–921