# TESD: A Novel Ground Truth Estimation Method

A. M. Biancardi and A. P. Reeves

School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA;

## ABSTRACT

Knowledge of the exact shape of a lesion, or ground truth (GT), is necessary for algorithm validation, measurement metric analysis, accurate size estimation. When multiple readers provide their documentations of a lesion that can ultimately be described with occupancy regions, estimating the unknown GT is achieved by aptly merging those occupancy regions into a single outcome. Several methods are already available but, even when they consider the spatial location of pixels, e.g. thresholded probability-map (TPM) or STAPLE, pixels are assumed spatially independent (even when STAPLE proposes a hidden-Markov-random-field fix). In this paper we propose Truth Estimate from Self Distances (TESD): a new voting scheme, for all the voxels inside and outside the occupancy region, in order to take in account three key characteristics: (a) critical shape conformations, like holes or spikes, that are defined by the reciprocally surrounding pixels, (b) marking co-locations, meaning the closeness without intersection of one reader's marking to other readers' ones and c) the three-dimensionality of lesions as imaged by CT scanners. In TESD each voxel is labeled into four categories according to its signed distance transform and then the labeled images are combined with a center of gravity method to provide the GT estimation. This theoretical approach was validated on a subset of the publicly available Lung Image Database Consortium archive, where a total of 35 nodules documented on 26 scans by all four radiologists were available. The results obtained are reasonable estimates, with GT obtained close to TPM and STAPLE; at the same time this method is not limited to the intersections of readers' marked regions.

**Keywords:** CAD development, ground-truth estimation, diagnosis, response to therapy, volumetric measurement

## 1. INTRODUCTION

The knowledge of the exact shape of the anatomical entities is becoming more and more important because of the continuous development of algorithms and systems that automate or assist radiologists in performing tasks based on image interpretation. For instance, in the case of lesions, the exact description of their three-dimensional extent is necessary for several tasks including algorithm validation, measurement metric analysis, and accurate size estimation. One possibility is to use phantoms that provide exact information on the goal being aimed at. The major criticism that has been moved against this kind of experiments is that the phantom is not complex enough to bring all of the real case variability and task difficulty because of, most of all, its being inanimate. This last objection holds true also on experiments conducted on dead bodies where the specimens were extracted for measuring after being imaged. Hence, the only source that is currently accepted for the estimation of real extents is what is perceived by expert radiologists and made available as documented scans . The problem here is dealing with readers' variability and more precisely dealing with how to find an estimate of the actual lesion from the set of documentations that, tough representing the same entity, show that readers' variability in its spatial extent On the other hand, the availability of multiple readers' markings makes it possible to generate an estimate that is expected to be a closer representation of the actual lesion region because it aims at minimizing the subjectivity of each reader's marking. Therefore, the way this estimate is generated takes an important role in the performance of the aforementioned goals (validation, metric analysis, ...).

Methods for the estimation of Ground Truths (GTs), based on the spatial processing of readers' markings, already exist, such as Thresholded Probability-Maps[1] ($TPM$) and Simultaneous Truth and Performance Level Estimation[2] ($STAPLE$), but, as discussed further on, they assume spatial independence of the voxels making up

Further author information: (Send correspondence to A.M.B.)
A.M.B.: E-mail: amb284@cornell.edu, Telephone: 1 607 254 8819, Fax: 1 607 255 9072

Medical Imaging 2009: Computer-Aided Diagnosis, edited by Nico Karssemeijer, Maryellen L. Giger
Proc. of SPIE Vol. 7260, 72603V · © 2009 SPIE · CCC code: 1605-7422/09/$18 · doi: 10.1117/12.813438

the marked regions for their main analysis. In this paper a GT estimation method, Truth Estimate from Self Distances ($TESD$), which actually takes into account and makes use of spatial contiguity and three-dimensionality of the nodule, is detailed.

## 2. TESD ESTIMATOR

$TESD$ is an algorithm that estimates the unknown ground truth (GT) of a lesion from a set of readers' markings. 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 signed 3D euclidean distance transform $D_i$, i.e. 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 weighted label maps $L_i$ where each voxel is labeled, with a strength coefficient, into four categories according to its signed distance-transform: inner core, inside border, outside border, and outer space (Figure 1 shows an example of a labeled voting map for a single reader's marked region);
- the labeled maps $L_i$ are then combined by a center of gravity method to produce a global label map that is then processed to provide the final binary estimate.

## 3. MATERIALS

A subset of the publicly available Lung Image Database Consortium archive[3] was used to evaluate $TESD$. The Lung Image Database Consortium[4,5] is a cooperative program, started by the National Institutes of Health in 2000, aiming at the creation of a large database of whole lung CT scans, documented by means of nodule locations or boundaries drawn by multiple radiologists, for the development and the evaluation of different CAD approaches. 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 ground truth, independently from the actual number of radiologists that supplied the initial boundaries. Instead a double reading process, performed by every radiologist, was

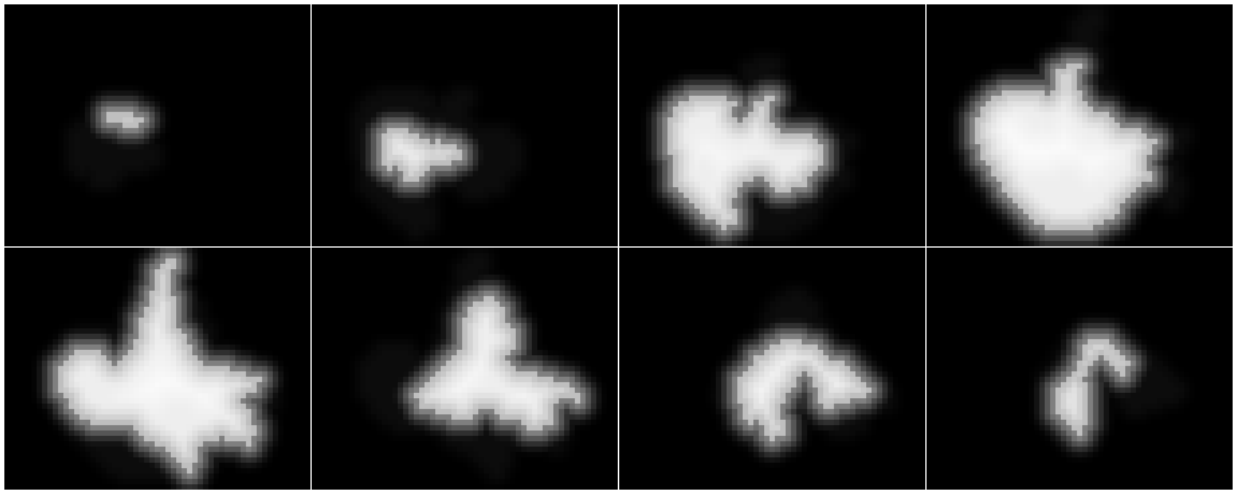

Figure 1. Tiled display of a labeled voting map for a reader-marked region: the brighter the voxel, the higher its value. The four label categories are easily perceivable.

established and drawings of nodule boundaries, in every axial image in which they appear, are provided for those nodules that were estimated to be 3 mm in size or bigger The availability of such annotated nodules makes the LIDC database an important source for the assessment of GT estimation methods.

For this paper 100 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).

## 4. METHODS

GT estimations by $TESD$, performed on the LIDC subset, are compared with the regions of two other GT estimating methods: $TPM_{0.5}$ and $STAPLE$. The $TPM_{0.5}$ estimator is computed by assigning to each voxel a value equal to average number of readers that included such voxel in their markings and then applying a threshold of 0.5 to give a Thresholded Probability-Map ($TPM_{0.5}$) that represent the regions marked by 2 or more readers. Our $STAPLE$ implementation is loosely based on a version from ITK,[6] for computing the EM-stage estimate, to which we added the max-flow min-cut optimization.[7]

The comparison is performed by determining the volumes of the GT estimates. Volume values are computed by counting the number of nodule pixels in each of the image slices and then multiplying their sum by the voxel volume;[8] this method is frequently used in CAD/CADx tools. Pixels belonging to any excluded inner regions are not counted when computing the nodule volume as they do not belong to the nodule region. When expressing the volumes in the uni-dimensional scale space, like the one used by RECIST,[9] the diameter $d$ of the equivalent sphere was used, i.e. the diameter of a sphere having the same volume as the estimate:

$$d = 2\sqrt[3]{\frac{3v}{4\pi}}$$

## 5. RESULTS

The GT volumes computed with the new method were in the range between -2.0% to +17.0% (average 5.1%, median 3.2%) with respect to $TPM_{0.5}$ estimates and in the range between -15.9% to 17.2% (average 2.4%, median 1.3%) with respect to $STAPLE$ estimates. When size is expressed as a uni-dimensional volumetric-based diameter, then the difference ranges were between -0.7% to 6.0% (average 1.8%, median 1.1%) for $TPM_{0.5}$ and between -5.0% to 6.1% (average 0.9%, median 0.4%) for $STAPLE$. Figure 2 shows an example of a GT estimate (j) by $TESD$, determined according to the readers' markings, (a) to (d), in the right column. Rows (e) to (h) show the weighted label maps computed from the three-dimensional distance transform of each reader's mark-up, while image (j) displays the global label map from which the final estimate is derived. Each set of tiles displays the data of an equivalent number of consecutive axial positions.

## 6. DISCUSSION

The motivations behind the development of a new GT estimation method lie in the need of giving equal weights to the readers that may disagree on the exact placement of the lesion border, but that largely agree on the lesion main core. This scenario translated into two requirements:

- the importance of not giving more reliability to one or more readers to the disadvantage of others, and
- the ability of perform 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 not adequate weighting labels that would depend on the shape of the marked region and 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. When comparing $TESD$ with the other methods, most of the differences arise from the use of a 3D distance transform:

1. $TESD$ can avoid the assumption of spatial independence of voxels,
2. the estimation is based on local information, and
3. the method is actually processing the volumes, making use of the images as a whole and not just as a collection of individual bi-dimensional slices.

As regards item 1, it is true that $STAPLE$ has a final optimization step that takes advantage of each voxel nearest neighbors to regularize the EM-stage estimation; however $STAPLE$ main estimation comes from the iterative EM stage and there the spatial independence of voxel is assumed to derive the formulas that estimate readers performance. Voxel independence is assumed by definition by probability-map-based methods.

The effect of all these factors can be perceived in Figure 3, where $TESD$ estimate of the nodule extent differs from the other two methods, especially in the bottom right corner of the rightmost tile. In the left column, tiles (a) through (d) show the readers' markings; on the right column, the GT estimates of (e) $TPM_{0.5}$, (f) $STAPLE$, and (g) $TESD$ are shown. The largest differences between $TESD$ and $STAPLE$, however, occur when two of the readers have almost identical markings. In one of these cases, shown in Figure 4, the two overlapping boundaries, (a) and (c), have a volume of 8888 $mm^3$ and a volumetric-based diameter of 5.50 mm, which are the exactly the same values of the $STAPLE$ estimate (e), whereas the $TPM_{0.5}$ and $TESD$ estimates, (f) and (g) respectively, have volumes of and volumetric-based diameters of 5.74 mm and 6.03 mm, respectively.

## 7. CONCLUSIONS

A new ground truth estimation algorithm was presented. It was shown that it is able to provide reasonable estimates with results close to $TPM_{0.5}$ and $STAPLE$ and with its new use of the co-location of readers' markings, taking advantage of their proximity, and its utilization of the inherent three-dimensionality of the input data.

## ACKNOWLEDGMENTS

This research was supported in part by NIH grant R33CA101110.

## REFERENCES

[1] Meyer, C. R., Johnson, T. D., McLennan, G., Aberle, D. R., Kazerooni, E. A., MacMahon, H., Mullan, B. F., Yankelevitz, D. F., van Beek, E. J. R., Armato III, S. G., McNitt-Gray, M. F., Reeves, A. P., Gur, D., Henschke, C. I., Hoffman, E. A., Bland, P. H., Laderach, G., Pais, R., Qing, D., Piker, C., Guo, J., Starkey, A., Max, D., Croft, B. Y., and Clarke, L. P., "Evaluation of lung MDCT nodule annotation across radiologists and methods," *Academic Radiology* **13**, 1254–1265 (2006).

[2] Warfield, S., Zou, K., and Wells, W., "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *Medical Imaging, IEEE Transactions on* **23**(7), 903–921 (July 2004).

[3] National Cancer Institute, "National cancer imaging archive." https://imaging.nci.nih.gov/ncia/. Accessed Jan 12, 2009.

[4] 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 Jan 12, 2009.

[5] National Cancer Institute, "Lung imaging database consortium (LIDC)." http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC. Accessed Jan 12, 2009.

[6] Ibanez, L., Schroeder, W., Ng, L., and Cates, J., *The ITK Software Guide*. Kitware, Inc. ISBN 1-930934-15-7, http://www.itk.org/ItkSoftwareGuide.pdf, second ed. (2005).

[7] Greig, D. M., Porteous, B. T., and Seheult, A. H., "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society* **51**(2), 271–279 (1989).

[8] Breiman, R. S., Beck, J. W., Korobkin, M., Glenny, R., Akwari, O. E., Heaston, D. K., Moore, A. V., and Ram, P. C., "Volume determinations using computed tomography," *American Journal of Roentgenology* **138**(2), 329–333 (1982).

[9] Therasse, P., Arbuck, S., Eisenhauer, E., Wanders, J., Kaplan, R., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A., Christian, M., and Gwyther, S., "New guidelines to evaluate the response to treatment in solid tumors," *J. Natl. Cancer Inst.* **92**, 205–216 (February 2000).

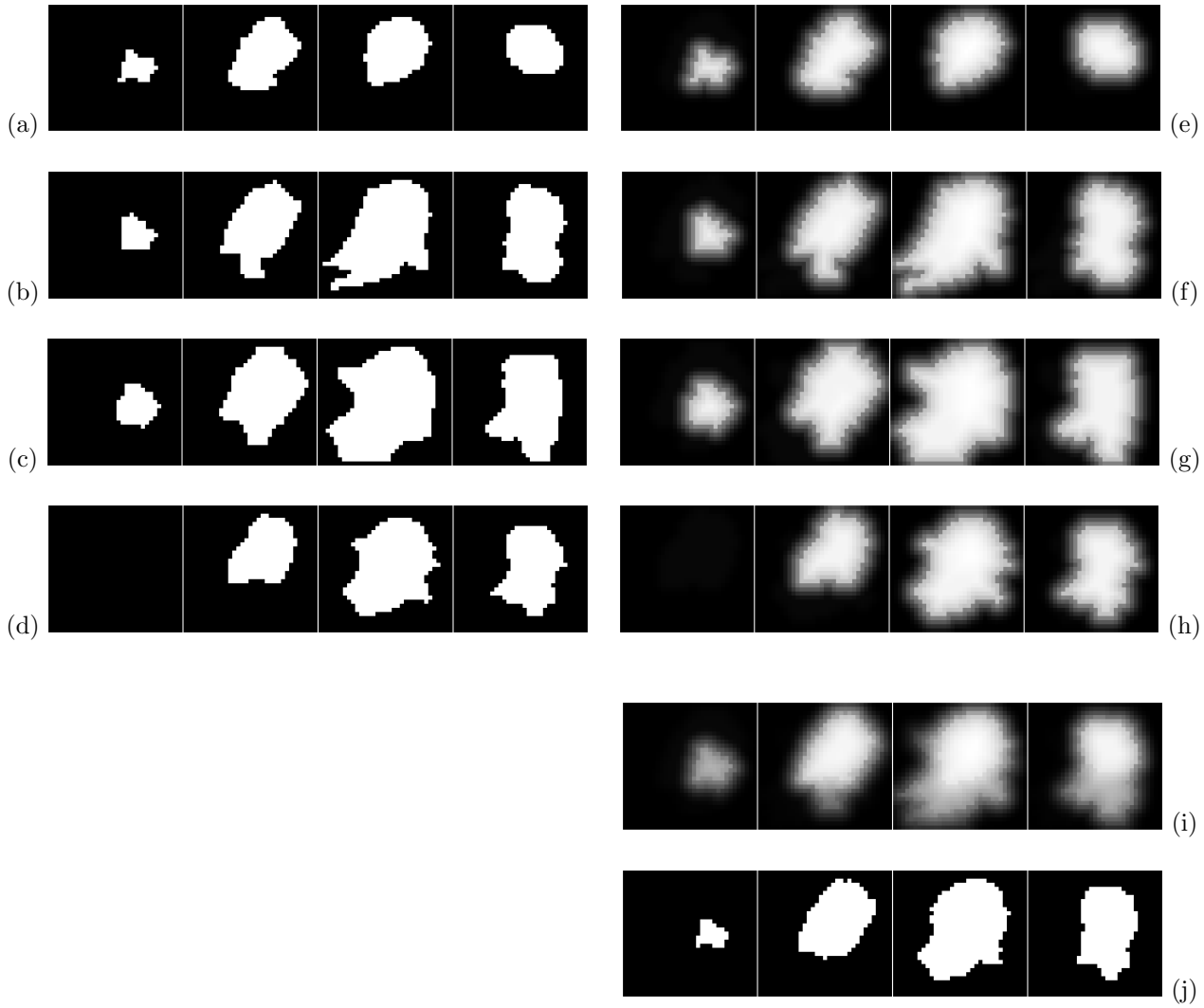


Figure 2. An example of GT estimations. On the left column tiles (a) to (d) show the readers' markings; on the right column, rows (e) to(h) display the respective weighted labeled maps for the readers' markings on their right, (i) is the global label map and (j) is $TESD$ GT estimate. Each set of tiles displays the data of an equivalent number of consecutive axial positions.

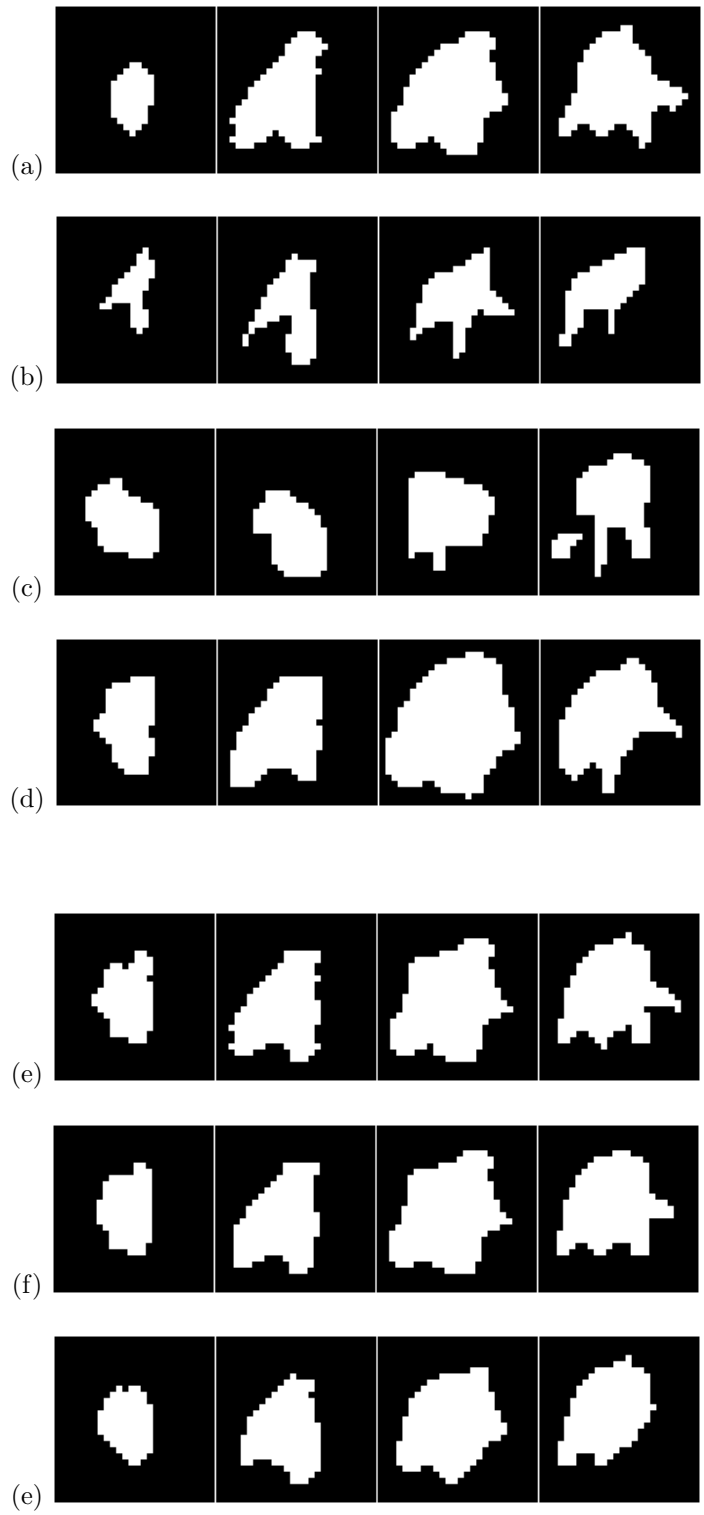


Figure 3. An example of nodule where $TESD$ estimate differs from the other two methods. On the left column tiles (a) to (d) show the readers' markings; on the right column, the GT estimates of (e) $TPM_{0.5}$, (f) $STAPLE$, and (g) $TESD$ are shown

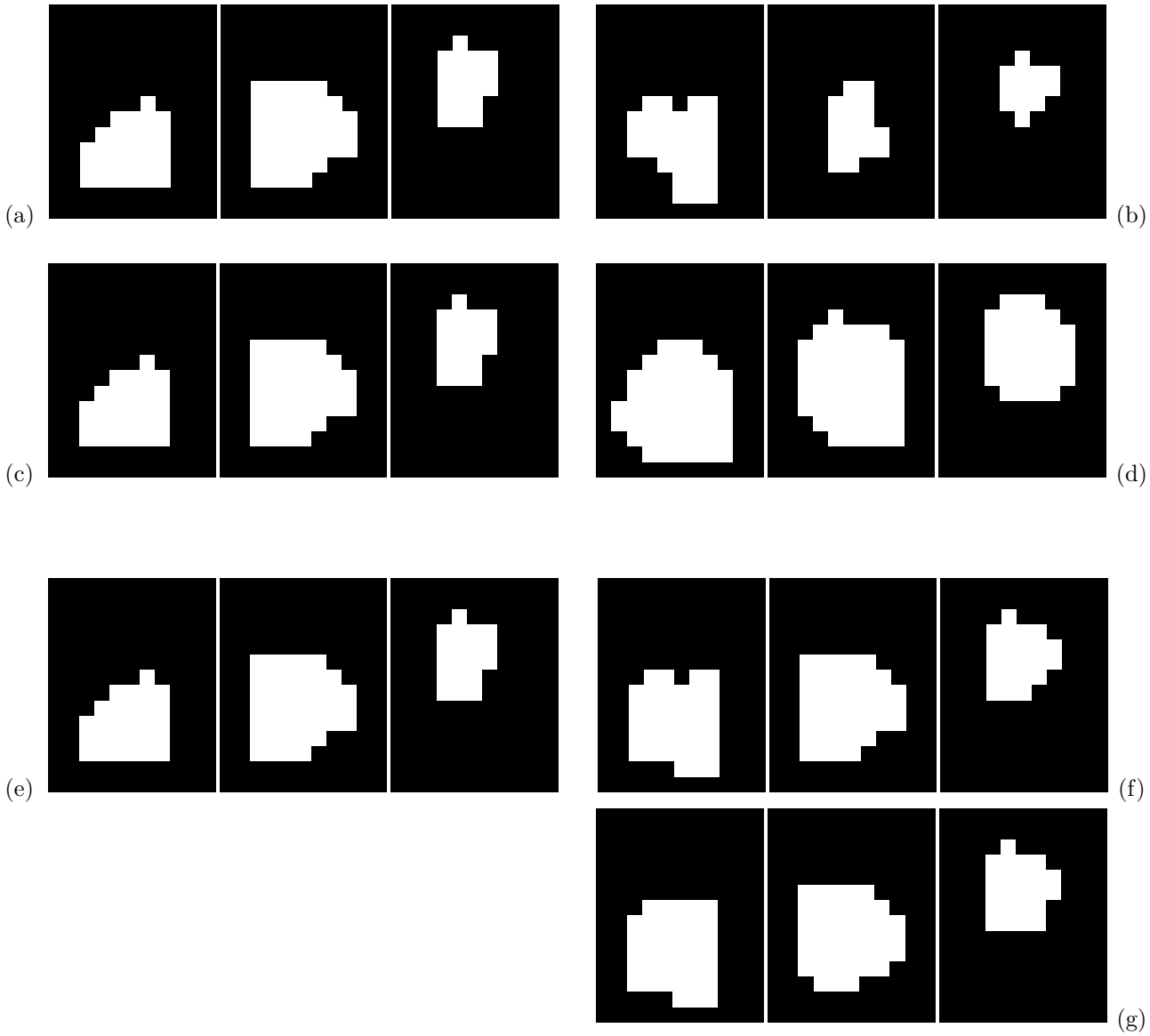


Figure 4. An example of nodule where two overlapping boundaries are present. In the top part, tiles (a) to (d) show the readers' markings; in the bottom part, the GT estimates of (e) $STAPLE$, (f) $TPM_{0.5}$, and (g) $TESD$ are shown