# An Analysis of Two Ground Truth Estimation Methods

A. M. Biancardi,[a] A. C. Jirapatnakul,[a] S. Fotin,[a] T. Apanasovich,[b] and A. P. Reeves,[a]

[a]Electrical and Computer Engr., Cornell University, Ithaca, NY, USA;
[b]JMC, Thomas Jefferson University

## ABSTRACT

An estimation of the so called Ground Truth (GT), i.e. the actual lesion region, can minimize readers' subjectivity if multiple readers' markings are combined. Two methods perform this estimate by considering the spatial location of voxels: Thresholded Probability-Map (TPM) and Simultaneous Truth and Performance Level Estimation (STAPLE). An analysis of these two methods has already been performed. The purpose of this study, however, is gaining a new insight into the method outcomes by comparing the estimated regions. A subset of the publicly available Lung Image Database Consortium archive was used, selecting pulmonary nodules documented by all four radiologists. The TPM estimator was 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. Our STAPLE implementation is loosely based on a version from ITK, to which we added the graph cut post-processing. The pair-wise similarities between the estimated ground truths were analyzed by computing the respective Jaccard coefficients. Then, the sign test of the differences between the volumes of TPM and STAPLE was performed. A total of 35 nodules documented on 26 scans by all four radiologists were available. The spatial agreement had a one-sided 90% Confidence Interval of [0.92, 1.00]. The sign test of the differences had a p-value less than 0.001. We found that (a) the differences in their volume estimates are statistically significant, (b) the spatial disagreement between the two estimators is almost completely due to the exclusion of voxels marked by exactly two readers, (c) STAPLE tends to weight more, in its GT estimate, readers marking broader regions.

**Keywords:** CAD development, algorithm validation, diagnosis, response to therapy, volumetric 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, 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.[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.

The Lung Image Database Consortium[9] is one of the answers to this need. This consortium 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 multiple radiologists without a consensus stage, for the development and the evaluation of different CAD approaches. The availability of such annotated nodules makes the LIDC database an invaluable source also 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

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, 72600E · © 2009 SPIE · CCC code: 1605-7422/09/$18 · doi: 10.1117/12.813440

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. In this paper two methods to estimate GTs are analyzed: thresholding at 0.5 of probability maps ($TPM_{0.5}$) and simultaneous truth and performance estimation ($STAPLE$). Previously[10] a separate analysis of $TPM_{0.5}$s and $STAPLE$ (without the graph-cut post-processing) was conducted; here, a detailed comparison of the GT regions created by $TPM_{0.5}$ and $STAPLE$ is performed.

## 2. GROUND TRUTH ESTIMATORS

There are several approaches for the estimation of ground truth from several expert readers. In this study, two of the most used approaches that perform this estimate by considering the spatial location of voxels were considered: Thresholded Probability-Maps ($TPM$) and Simultaneous Truth and Performance Level Estimation ($STAPLE$). In a Probability-Map,[1] 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 1 by 3 out of the 4 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, voxels present in none of the segmentations have a value of 0, and voxels in some of the segmentations but not all will have a value of 0.25, 0.50, or 0.75. To generate an estimated ground truth, the Probability-Map may be thresholded at a particular value. In this study, the Probability-Maps are thresholded at 0.50 to give a Thresholded Probability-Map ($TPM_{0.5}$) that represent the regions marked by 2 or more readers.

The second approach considered in this study is an algorithm proposed by Warfield et al.,[11] simultaneous truth and performance estimation ($STAPLE$). In this method, the true ground truth is treated as a hidden variable and therefore not directly observable. Since the true ground truth is unknown, reader performance is also unknown. This method estimates both the ground truth and reader performance simultaneously using an expectation-maximization (EM) algorithm. The results of the EM stage 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 ground truth, 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)[12] 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 ground truth. Following the authors' indications,[11] 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 ground truth segmentation. A reasonable value to use is the relative proportion of pixels marked as belonging to the input segmentations[11] 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 was realized as a max-flow min-cut, as suggested in[11], by a program that we developed, based on the analysis by Boykov and Kolmogorov[13] of the algorithm formulated by Ford and Fulkerson.[14]

## 3. MATERIALS AND METHODS

The comparison was performed on whole-lung CT scans provided by the LIDC archive.[15] The LIDC process model[16,17] 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. 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 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.

After determining the GT estimates for all of the three 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;[18] this method is frequently used in CAD/CADx tools. When expressing the volumes in the uni-dimensional scale space, like the one used by RECIST,[19] the diameter $d$ of the equivalent sphere was used, i.e. the diameter of a sphere having the same volume as the estimate: $d = 2\sqrt[3]{\frac{3v}{4\pi}}$ Then two indicators of the pair-wise similarities between those estimates were computed: the Fisher sign test of the differences between the respective estimate volumes and one-sided 90% confidence interval (CI) of the Jaccard coefficients between the respective estimate 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 differences to follow the binomial distribution with probability $p = 0.5$. Given two sets $X$ and $Y$, the Jaccard coefficient[20] 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 $TPM_{0.5}$ 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 (0.5).

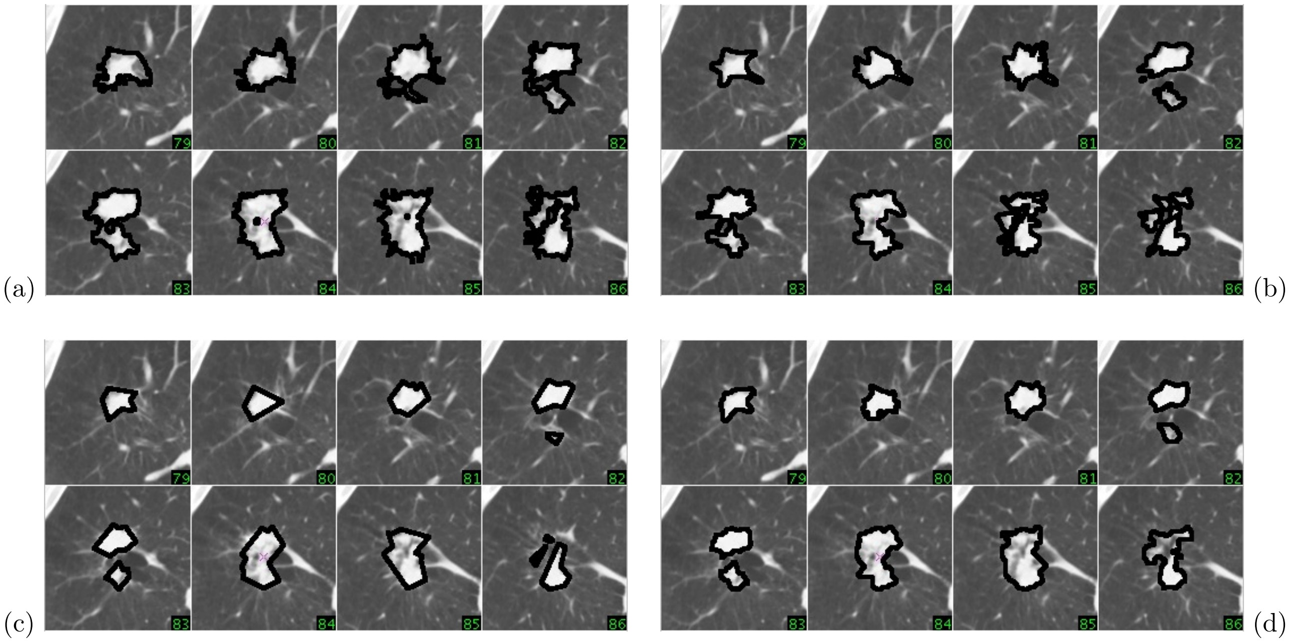

Figure 1. An example of nodule documentation provided by the LIDC database. In this case a montage of the central slices of a nodule are displayed and, overlaid, the radiologists' markings of the nodule boundary. Each montage set (a), (b), (c), and (d) shows the respective central slice portion of the mark-up of one radiologist.

It is important to underline that all the GT estimations were performed considering each nodule separately without being able to extend the reader performance level evaluation to more than a single nodule because no reader can be tracked among the full set of nodules or even within the same case. This is due to the fact that all the LIDC data is anonymized, i.e. not only the DICOM image data, but also all the documentation attached to each scan, and 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[17]);
- 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.

## 4. RESULTS

A total of 35 nodules documented by all four radiologists were selected. The median diameters from the manual markings ranged from 4.4 to 23.8 mm (mean 11.1 mm, median 7.46 mm). The spatial agreement between the ground truth estimates of the $TPM_{0.5}$ and $STAPLE$ had a one-sided 90% C.I. of $[0.92, 1.00]$: this was reflected into relative volume differences ranging from -0.3% to 12.3% (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 0.44%). The sign test on the difference between the volume estimations had a p-value less than 0.001. As regards the differences between the intermediate EM outputs and the final $STAPLE$ regions, the spatial agreement had a one-sided 90% C.I. of $[0.96, 1.00]$; relative volume differences ranged between -0.3% to 7.9% (average 1.6%, median 0.7%) or, expressing them in a uni-dimensional scale space, into a range from -0.1% to 2.7% (average 0.5%, median 0.2%). The sign test on the difference between the volume estimations had a p-value less than 0.001.

## 5. DISCUSSION

The Jaccard coefficients showed a certain degree of disagreement between the two methods, the one-sided sign test on the volume differences showed that $STAPLE$ volumes are almost always smaller than $TPM_{0.5}$ 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 $TPM_{0.5}$ estimate and includes 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 $TPM_{0.5}$ 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 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 $TPM_{0.5}$ 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 $TPM_{0.5}$ and $STAPLE$ concerns each reader' reliability: the $TPM_{0.5}$ assumes 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 this feature tries to capture readers' agreements and disagreements, it may also be affecting the results. In those 6 case, brought to our attention by the EM result analysis, $STAPLE$ shows the following behavior: (a) when there are two readers' markings almost coinciding, 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 Figure 2; (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 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 2 shows an example when there are two almost completely overlapping boundaries (b) and (c): in this case $STAPLE$ estimate (f) is determined by that region, while the $TPM_{0.5}$ GT estimate includes voxels marked by the other readers. Figure 3, on the other hand, shows a case where $STAPLE$ and $TPM_{0.5}$ GT estimates differ thanks to the geometric regularization of the estimate, in $STAPLE$, owing to the use of the hidden Markov random field. Final estimates are shown in (f), $TPM_{0.5}$, and (h), $STAPLE$, while intermediate outcomes for the two methods are shown in (e), the probability map before thresholding, and in (g), the output of the EM stage.

## 6. CONCLUSIONS

A comparison of the GT regions created by $TPM_{0.5}$ and $STAPLE$ was performed. We found that (a) the differences in the volume estimates of $TPM_{0.5}$ and $STAPLE$ are statistically significant being the volumes of the $TPM_{0.5}$ estimated GT almost always grater than $STAPLE$ ones, (b) the spatial disagreement between the two estimators is almost completely due to the final regularization stage of $STAPLE$ that mostly excludes voxels marked by exactly two readers, (c) $STAPLE$ results can be affected either by the closeness between two mark-ups or by the larger regions being favored over smaller ones.

## 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] Reeves, A. P., Biancardi, A. M., Apanasovich, T. V., Meyer, C. R., MacMahon, H., van Beek, E. J., Kazerooni, E. A., Yankelevitz, D., McNitt-Gray, M. F., McLennan, G., Armato III, S. G., Henschke, C. I., Aberle, D. R., Croft, B. Y., and Clarke, L. P., "The lung image database consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements," *Academic Radiology* **14**, 1475–1485 (Dec 2007).

[3] Ko, J. P., Rusinek, H., Jacobs, E. L., Babb, J. S., Betke, M., McGuinness, G., and Naidich, D. P., "Small pulmonary nodules: Volume measurement at chest CT – phantom study," *Radiology* **228**, 864–870 (September 2003).

[4] Kuhnigk, J.-M., Dicken, V., Bornemann, L., Wormanns, D., Krass, S., and Peitgen, H.-O., "Fast automated segmentation and reproducible volumetry of pulmonary metastases in CT-scans for therapy monitoring," in [*Lecture Notes in Computer Science*], **3217**, 933–941, Medical Image Computing and Computer-Assisted Intervention, Springer-Verlag GmbH (2004).

[5] Okada, K., Comaniciu, D., and Krishnan, A., "Robust anisotropic gaussian fitting for volumetric characterization of pulmonary nodules in multislice CT," *IEEE Transactions on Medical Imaging* **24**, 409–423 (March 2005).

[6] Goodman, L. R., Gulsun, M., Washington, L., Nagy, P. G., and Piacsek, K. L., "Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements," *American Journal of Roentgenology* **186**, 989–994 (April 2006).

[7] Revel, M.-P., Merlin, A., Peyrard, S., Triki, R., Couchon, S., Chatellier, G., and Frija, G., "Software volumetric evaluation of doubling times for differentiating benign versus malignant pulmonary nodules," *American Journal of Roentgenology* **187**, 135–142 (July 2006).

[8] Reeves, A., Chan, A., Yankelevitz, D., Henschke, C., Kressler, B., and Kostis, W., "On measuring the change in size of pulmonary nodules," *IEEE Transactions on Medical Imaging* **25**, 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 Jan 9, 2009.

[10] Ross, J. C., Miller, J. V., Turner, W. D., and Kelliher, T. P., "An analysis of early studies released by the lung imaging database consortium (LIDC)," *Academic Radiology* **14**, 1382–1388 (Nov 2007).

[11] 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).

[12] 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).

[13] Boykov, Y. and Kolmogorov, V., "An experimental comparison of Min-Cut/Max-Flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 1124–1137 (September 2004).

[14] Ford Jr, L. R. and Fulkerson, D. R., "Maximal flow through a network," *Canadian Journal of Mathematics* **8**, 399–404 (June 1956).

[15] National Cancer Institute, "National cancer imaging archive." https://imaging.nci.nih.gov/ncia/. Accessed Jan 9, 2009.

[16] Armato, S. G., McLennan, G., McNitt-Gray, M. F., Meyer, C. R., Yankelevitz, D., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E., MacMahon, H., Reeves, A. P., Croft, B. Y., Clarke, L. P., and 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 (2004).

[17] McNitt-Gray, M. F., Armato III, S. G., Meyer, C. R., Reeves, A. P., McLennan, G., Pais, R. C., Freymann, J., Brown, M. S., Engelmann, R. M., Bland, P. H., Laderach, G. E., Piker, C., Guo, J., Towfic, Z., Qing, D. P.-Y., Yankelevitz, D. F., Aberle, D. R., van Beek, E. J., MacMahon, H., Kazerooni, E. A., Croft, B. Y., and Clarke, L. P., "The lung image database consortium (LIDC) data collection process for nodule detection and annotation," *Academic Radiology* **14**, 1464–1474 (Dec 2007).

[18] 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).

[19] 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).

[20] Jaccard, P., "Nouvelles recherches sur la distribution florale," *Bull. Soc. Vaudoise Sci. Nat* (44), 223–270 (1908).

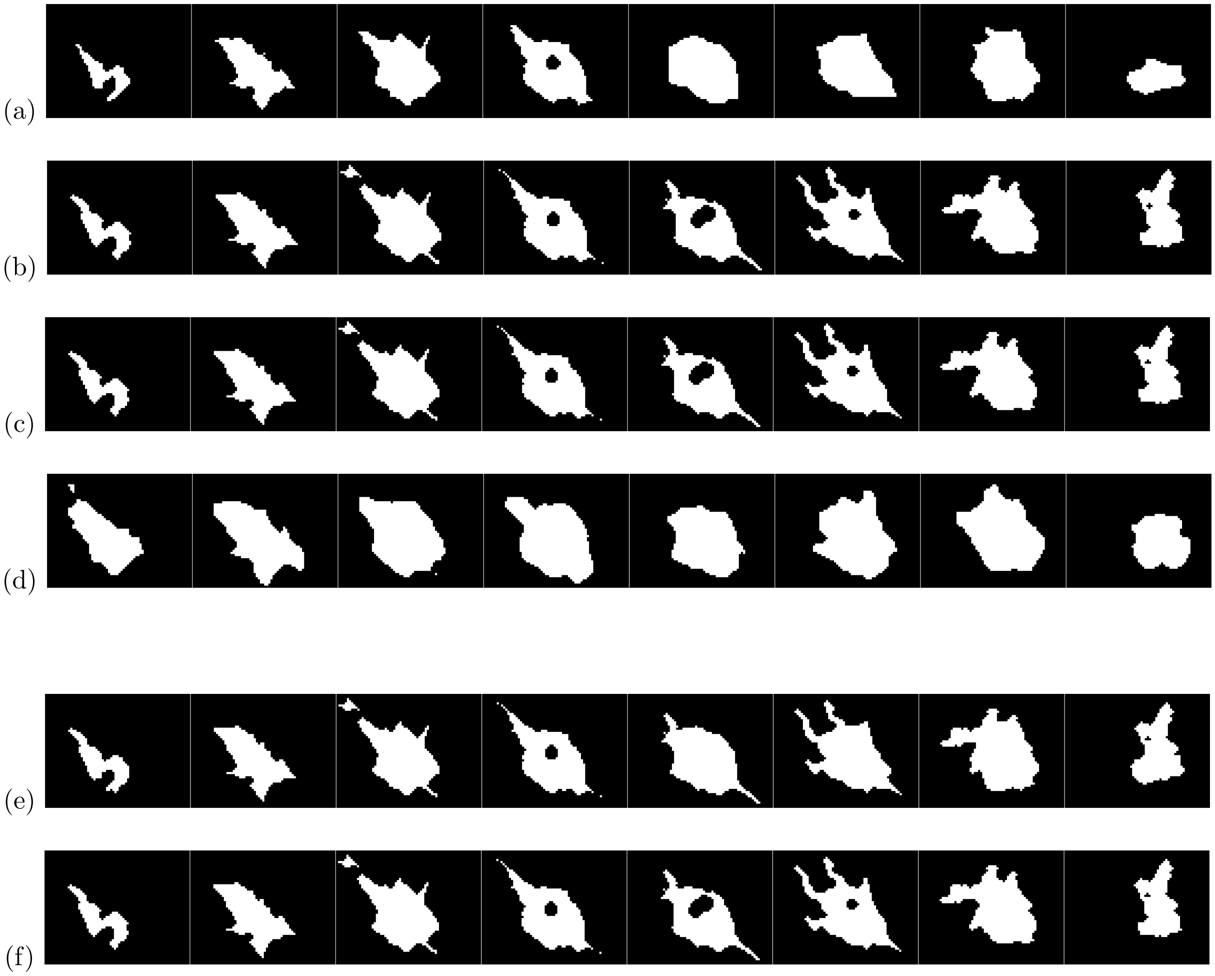

Figure 2. An example of nodule where two overlapping boundaries, (b) and (c), determine $STAPLE$ estimate (f). Reader marked regions are displayed in the top part, (a) to (d), while (e) is the $TPM_{0.5}$ estimate.

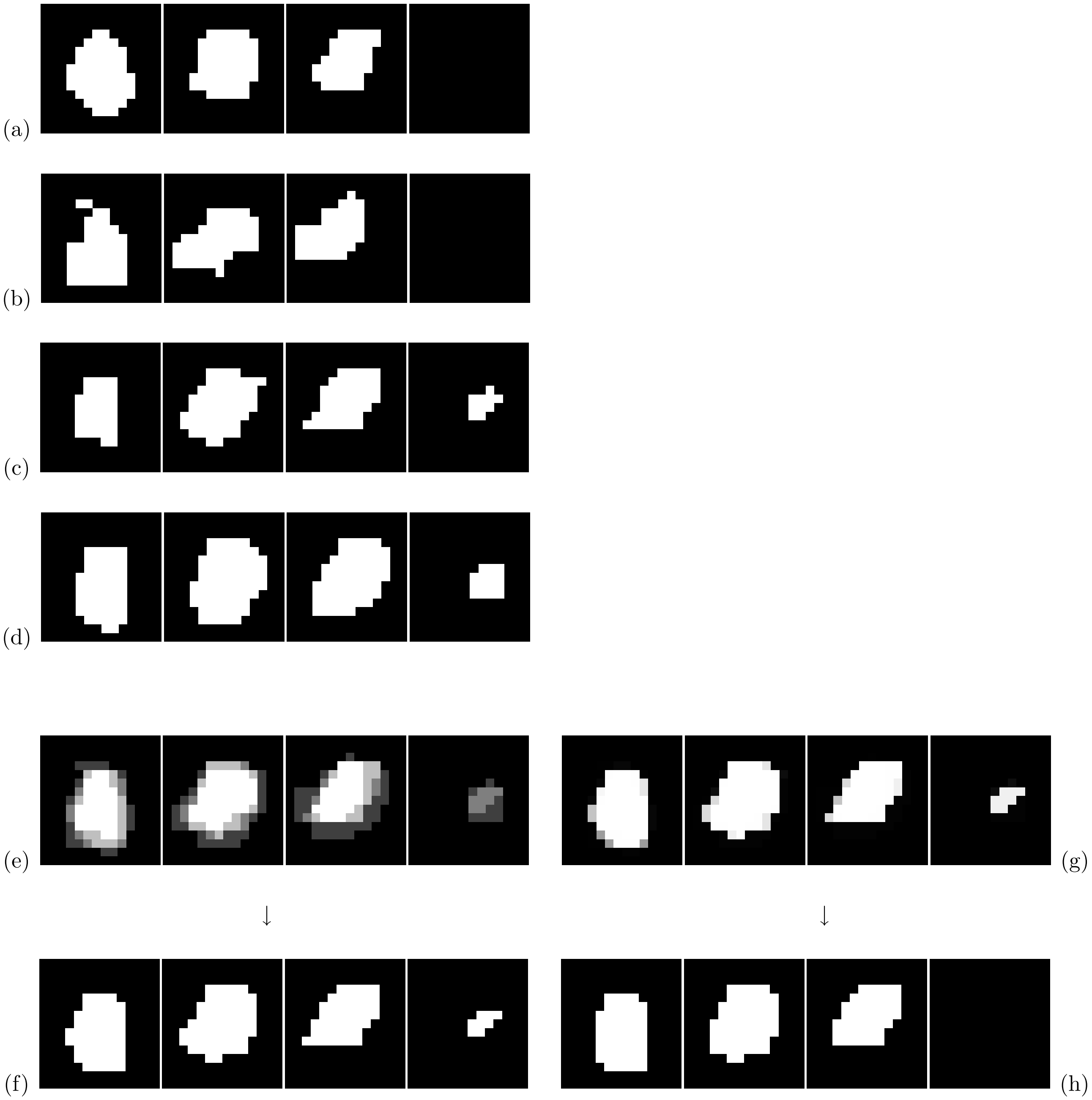

Figure 3. An example of nodule where $STAPLE$ and $TPM_{0.5}$ estimates differ because of the regularization of the estimate, in $STAPLE$, thanks to the use of the hidden Markov random field. Final estimates are shown in (f), $TPM_{0.5}$, and (h), $STAPLE$, while intermediate outcomes for the two methods are shown in (e), the probability map before thresholding, and in (g), the output of the EM stage.