Estimation of Inferential Uncertainty in Assessing Expert Segmentation Performance from STAPLE

Olivier Commowick[1] and Simon K. Warfield[1]

Computational Radiology Laboratory, Department of Radiology,
Children's Hospital, 300 Longwood Avenue, Boston, MA, 02115, USA
{Olivier.Commowick,Simon.Warfield}@childrens.harvard.edu

Abstract. The evaluation of the quality of segmentations of an image, and the assessment of intra- and inter-expert variability in segmentation performance, has long been recognized as a difficult task. Recently an Expectation Maximization (EM) algorithm for Simultaneous Truth and Performance Level Estimation (STAPLE), was developed to compute both an estimate of the reference standard segmentation and performance parameters from a set of segmentations of an image. The performance is characterized by the rate of detection of each segmentation label by each expert in comparison to the estimated reference standard.

This previous work provides estimates of performance parameters, but does not provide any information regarding their uncertainty. An estimate of this inferential uncertainty, if available, would allow estimation of confidence intervals for the values of the parameters, aid in the interpretation of the performance of segmentation generators, and help determine if sufficient data size and number of segmentations have been obtained to accurately characterize the performance parameters.

We present a new algorithm to estimate the inferential uncertainty of the performance parameters for binary segmentations. It is derived for the special case of the STAPLE algorithm based on established theory for general purpose covariance matrix estimation for EM algorithms. The bounds on performance estimates are estimated by the computation of the observed Information Matrix. We use this algorithm to study the bounds on performance estimates from simulated images with specified performance parameters, and from interactive segmentations of neonatal brain MRIs. We demonstrate that confidence intervals for expert segmentation performance parameters can be estimated with our algorithm. We investigate the influence of the number of experts and of the image size on these bounds, showing that it is possible to determine the number of image segmentations and the size of images necessary to achieve a chosen level of accuracy in segmentation performance assessment.

1 Introduction

The evaluation of image segmentation has long been recognized as a difficult problem. Many methods have been proposed in the literature to deal with it. These can be classified into two groups. First, the evaluation can be based on distances between surfaces extracted from the automatic and the manual segmentation. For example, these can be the Hausdorff distance [1] or a mean distance between the two surfaces [2]. The other class of measures are voxel-based measures, i.e. overlap measures based on voxel-wise computations. Among those, the Dice similarity coefficient [3] or the Jaccard similarity coefficient [4, 5] have been widely used to measure the overlap between two segmentations.

These two classes of measures have their advantages and drawbacks, and both may be used to provide insight into the quality of a segmentation [6] and to compare segmentations. The evaluation of different experts or algorithms for a particular task can be done quantitatively, with performance characterized by rates of detection of labels, when a reference standard segmentation is available.

Segmentation performance characterization can also be achieved when no external reference standard segmentation is available by estimating the reference standard. One algorithm for this, called Staple [7], uses an Expectation-Maximization (EM) algorithm to estimate iteratively, from a set of N expert segmentations, the hidden reference standard segmentation and performance parameters for each segmentation. These parameters characterize the agreement of a given expert with the underlying reference standard.

The Staple algorithm generates only point estimates of the performance parameters, and provides no information about the uncertainty in the values of the parameters. Precise knowledge of the inferential uncertainty would enhance our ability to interpret the performance of segmentation generators, and could be used to determine if sufficient data size and number of segmentations have been obtained to accurately characterize the performance parameters. An estimate of this inferential uncertainty, if available, would describe confidence intervals for the values of the parameters. Such a confidence interval describes the certainty with which we know the value of the parameter. A different concept is the confidence interval for rater performance, which describes the range of performance we expect to see across repeated segmentations by the same rater. If the inferential uncertainty of the values of performance parameter estimates are very small, then a confidence interval for rater performance can be estimated simply by the sample variance over repeated segmentations.

We describe here an algorithm to estimate the inferential uncertainty of the performance parameters. We demonstrate this can be achieved by estimating the covariance matrix of the performance parameters from Staple by calculation of the observed Information Matrix. The computation of the observed Information Matrix has been described in the general EM framework [8]. In this paper we derive analytic closed form expressions necessary to compute the covariance of the performance parameters obtained from Staple in the case of binary segmentations. We then demonstrate factors that influence the uncertainty in the estimated performance parameters with simulated segmentations of images, and apply our algorithm to characterize the segmentation of unmyelinated white matter from MRI of brains of newborn infants.

2 Method

2.1 The Staple Algorithm

We first recall briefly the principle of the Staple algorithm [7]. This method uses as an input a set of segmentations from J experts (either manual delineations or automatic segmentations). These segmentations are available as decisions d_{ij}, indicating the label given by each expert j for each voxel i. The goal of Staple is to estimate both the reference segmentation $\mathbf{T}$ underlying the expert segmentations, and parameters $\theta = \{\theta_1, \ldots, \theta_j, \ldots, \theta_J\}$ describing the agreement between the experts and the hidden reference standard. In the general case, each of the parameters θ_j is an $L \times L$ matrix, where L is the number of labels in the segmentation, and $\theta_{js's}$ is the probability that the expert j gave the label s' to a voxel i instead of the label s, i.e. $\theta_{js's} = P(d_{ij} = s' | T_i = s)$.

If the reference standard was known, then estimating the performance parameters for each expert would be straightforward. However, as it is unknown, an EM approach [9, 8] is used to estimate the reference standard $\mathbf{T}$ and the performance parameters of the experts. The EM algorithm proceeds iteratively, alternating two steps:

- E-Step: Compute the expected value of the complete data log-likelihood $Q(\theta|\theta^{(k)})$ knowing the expert parameters at the preceding iteration: $\theta^{(k)}$. Evaluating this expression requires the knowledge of the posterior probability of the true score T: $P(T|D, \theta^{(k)})$, which is sufficient in this case to perform the Maximization step.
- M-Step: Estimate the performance parameters at iteration $k+1$, $\theta^{(k+1)}$ by maximizing the expected complete data log-likelihood $Q(\theta|\theta^{(k)})$, knowing the current estimate of the reference standard.

2.2 Covariance and Information Matrix

We are interested in the computation of the covariance matrix $C(\theta)$ of the expert parameters obtained by the Staple algorithm. This is done via the computation of the observed Information Matrix $I(\theta)$ of the parameters obtained after convergence of the EM algorithm. Then, the covariance matrix is obtained using the well-known result [10]: $C(\theta) = I^{-1}(\theta)$.

If all the data was known, the Information Matrix would be simply the matrix of the second derivatives of the log-likelihood function. However, in the case of an EM algorithm such as Staple, the hidden variables are unknown and their value may only be estimated. As some variables are hidden, only the observed Information Matrix $I(\theta)$ can be computed. The expression of $I(\theta)$ has been derived for a general EM algorithm in [8] (page 100).

We proceed by first computing the expected complete data Information Matrix $I_c(\theta)$ using the expected complete data log-likelihood $Q(\theta|\theta^{(k)})$ estimated in the EM algorithm. Then, to account for the uncertainty from the missing data, the expected missing data Information Matrix $I_m(\theta)$ is subtracted from $I_c(\theta)$ to obtain the observed Information Matrix, i.e. $I(\theta) = I_c(\theta) - I_m(\theta)$.

2.3 Computation of the Observed Information Matrix

We derive here the expression of the observed Information Matrix for the STAPLE algorithm in the binary case. In this case, each expert has delineated one structure by attributing the value 1 to a voxel belonging to the structure and 0 otherwise (background). In this particular case, the θ parameters can be represented entirely by two parameters for each expert j: $p_j = P(d_{ij} = 1|T_i = 1)$ and $q_j = P(d_{ij} = 0|T_i = 0)$. p_j is also known as the sensitivity of the expert j while q_j is also known as the specificity. To simplify as much as possible the notation for the following equations, we use the general notation $\theta_{js's}$ for the performance parameters, keeping in mind that only $p_j = \theta_{j11}$ and $q_j = \theta_{j00}$ are the meaningful parameters (θ_{j01} and θ_{j10} being completely determined as $\theta_{j01} = 1 - p_j$ and $\theta_{j10} = 1 - q_j$). Then, the EM algorithm is used to compute iteratively the expected value of the complete data log-likelihood function $Q(\theta|\theta^{(k)})$:

$$Q(\theta|\theta^{(k)}) = \sum_j \sum_i \left(W_i^{(k)} \log(\theta_{j,d_{ij},1}) + (1 - W_i^{(k)}) \log(\theta_{j,d_{ij},0}) \right) \tag{1}$$

where $\theta_{j,d_{ij},s}$ corresponds to either θ_{j0s} or θ_{j1s} depending on the decision d_{ij}. $W_i^{(k)}$ corresponds to the probabilistic estimate at the voxel i and the iteration k of the reference standard segmentation. Using this function, we now derive the observed Information Matrix of the parameters θ.

Derivation of the Expected Complete Data Information Matrix This matrix, denoted $I_c(\theta)$, is expressed as the second derivatives of the expected value of the complete data log-likelihood function [8], i.e.

$$\mathbf{I}_c(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T} Q(\theta|\theta^{(k)}) \tag{2}$$

Eq. (1) and Eq. (2) demonstrate that the non-diagonal terms of I_c are zero as the parameters are independent of each other. Therefore, I_c is a diagonal matrix composed of the following terms:

$$\mathbf{I}_{c;p_j} = \sum_i \frac{W_i^{(k)}}{\theta_{j,d_{ij},1}^2} \tag{3}$$

$$\mathbf{I}_{c;q_j} = \sum_i \frac{1 - W_i^{(k)}}{\theta_{j,d_{ij},0}^2} \tag{4}$$

Derivation of the Expected Missing Data Information Matrix Once I_c has been computed, the observed Information Matrix is obtained by subtracting from it the expected missing data Information Matrix I_m. This matrix is generally more difficult to compute than $I_c(\theta)$. When no analytical expression can be derived, it can be estimated using the EM algorithm itself to compute the

Jacobian matrix via numerical differentiation (see [10, 8]). In the general case of any EM algorithm, an analytic expression of I_m may also be obtained by the following equation [11] if the required derivatives exist:

$$\mathbf{I}_m(\theta) = \frac{\partial^2 Q(\theta|\theta^{(k)})}{\partial\theta^{(k)}\partial\theta^T} \tag{5}$$

In the case of the Staple algorithm, the expected value of the complete data log-likelihood function $Q(\theta|\theta^{(k)})$ can be differentiated. We have therefore derived the analytic expression of I_m elements as follows:

$$\frac{\partial^2 Q}{\partial\theta_{jtt}\partial\theta_n^{(k)}} = \sum_i \frac{(-1)^{1+d_{ij}}}{\theta_{j,d_{ij},t}} \frac{\partial W_i^{(k)}}{\partial\theta_n^{(k)}} \tag{6}$$

where t is either 1 or 0, to derive the expressions for p_j and q_j. Interestingly, it can also be shown that the obtained I_m matrix is symmetric, therefore minimizing the number of computations required. This expression gives I_m as a function of the derivatives of the probabilistic ground truth $W_i^{(k)}$. These $W_i^{(k)}$ have been derived by Warfield et al. [7] as:

$$W_i^{(k)} = \frac{f(T_i = 1)\prod_j \theta_{j,d_{ij},1}^{(k)}}{\sum_{m=0}^{1}\left(f(T_i = m)\prod_j \theta_{j,d_{ij},m}^{(k)}\right)} \tag{7}$$

For simplicity of notations, we will consider that the prior probability $f(T_i = 1)$, respectively $f(T_i = 0)$, is constant over the entire image and will abbreviate it by π_1, respectively π_0. However, all the derived expressions are still valid for spatially varying prior probabilities by replacing π_m in the following equations by $\pi_m(i)$. Knowing the expression of W_i, its derivative with respect to the expert parameters $p_n^{(k)} = \theta_{n11}^{(k)}$ and $q_n^{(k)} = \theta_{n00}^{(k)}$ can be derived:

$$\frac{\partial W_i^{(k)}}{\partial\theta_{ntt}^{(k)}} = (-1)^{1+d_{in}}\pi_0\pi_1 \frac{\left(\prod_{l\neq n}\theta_{l,d_{il},t}^{(k)}\right)\left(\prod_l \theta_{l,d_{il},1-t}^{(k)}\right)}{\left(\sum_{m=0}^{1}\pi_m\prod_l \theta_{l,d_{il},m}^{(k)}\right)^2} \tag{8}$$

where t is either 0 or 1. Therefore, $I_m(\theta)$, defined in Eq. (5), is computed by replacing $\frac{\partial W_i^{(k)}}{\partial\theta_n^{(k)}}$ by its value in Eq. (6). In practice, these values are computed easily by evaluating the different expressions at each voxel.

3 Results

To illustrate our formulation for deriving bounds on the value of the estimated segmentation parameters, we will present two applications. First, we show results with a simulated database, with specified parameters. Then, we present the application of our framework to provide insight into the confidence of the estimated parameters on a manually segmented neonate database.

3.1 Simulated Experiments

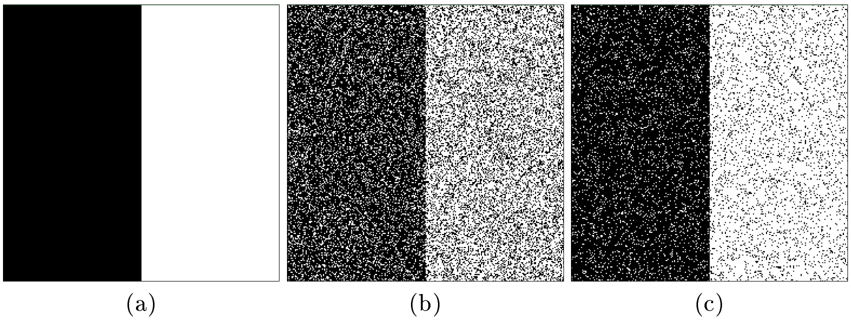


Fig. 1. Simulated Image Database. Simulated images used for the validation of our bounds estimation method : (a): known reference standard, (b): example of simulated segmentation of group 1 (sensitivity: 0.7, specificity: 0.8), (c): example of simulated segmentation of group 2 (sensitivity: 0.9, specificity: 0.9).

To evaluate our framework with respect to a known ground truth, we created a database of ten segmentations (image size 256 × 256), illustrated in Fig. 1, divided into two groups. From the ground truth in Fig. 1(a), we have therefore simulated a first group of 5 images with a sensitivity parameter of 0.7 and a specificity parameter of 0.8 (illustrated in image (b)). Then, a second group, illustrated in image (c), was generated with different parameters : sensitivity and specificity of 0.9. In order to test for the influence of the image size on the confidence in the parameters, we have also generated a second database with the same parameters but with image size of 128 × 128.

	256 × 256 Data		128 × 128 Data	
Segmentation #	Sens. (± StDev)	Spec. (± StDev)	Sens. (± StDev)	Spec. (± StDev)
1	0.7036 ± 0.0025	0.8011 ± 0.0022	0.6980 ± 0.0051	0.7988 ± 0.0045
2	0.7005 ± 0.0025	0.7964 ± 0.0022	0.6998 ± 0.0051	0.7960 ± 0.0045
3	0.7012 ± 0.0025	0.7909 ± 0.0023	0.6995 ± 0.0051	0.7980 ± 0.0045
4	0.6968 ± 0.0026	0.8003 ± 0.0022	0.6975 ± 0.0051	0.8007 ± 0.0044
5	0.7029 ± 0.0025	0.8010 ± 0.0022	0.6989 ± 0.0051	0.7964 ± 0.0045
6	0.9017 ± 0.0017	0.8973 ± 0.0017	0.8976 ± 0.0034	0.8998 ± 0.0034
7	0.9002 ± 0.0017	0.8995 ± 0.0017	0.8992 ± 0.0034	0.9012 ± 0.0034
8	0.8998 ± 0.0017	0.8986 ± 0.0017	0.9038 ± 0.0033	0.9027 ± 0.0033
9	0.8982 ± 0.0017	0.9018 ± 0.0017	0.8943 ± 0.0035	0.8960 ± 0.0034
10	0.8997 ± 0.0017	0.9007 ± 0.0017	0.8976 ± 0.0034	0.9036 ± 0.0033

Table 1. Simulated Evaluation of the Expert Parameters and their Bounds. Simulated experiments results showing the estimated parameters for each segmentation and its variability (one standard deviation). Results are shown for 256 × 256 images and 128 × 128 images, showing increased variability with decreasing image size.

We have then run Staple on those databases to estimate a reference standard and utilized our framework to estimate bounds on the estimated parameter values. The results are presented in Table 1 for the two databases. Our first observation on all our examples, including the following experiments on neonate data, was that the non diagonal terms of the covariance matrix were always much smaller than the diagonal terms. We have therefore chosen to present in this article the standard deviations obtained for each parameter, therefore neglecting the non-diagonal terms of the covariance matrix. The figures in Table 1 show that almost all the estimated parameters are correct, up to one standard deviation as estimated in our formulation. Deriving the bounds on these parameters therefore allows us to show that the estimation performed by Staple is accurate. The second observation that can be made on these figures is on the influence of the image size on the variability of the parameters. Our experiments indeed show a clear correlation between the image size and the variability, the standard deviations increasing when the image is subsampled.

3.2 Evaluation of Variability Parameters on a Neonate Database

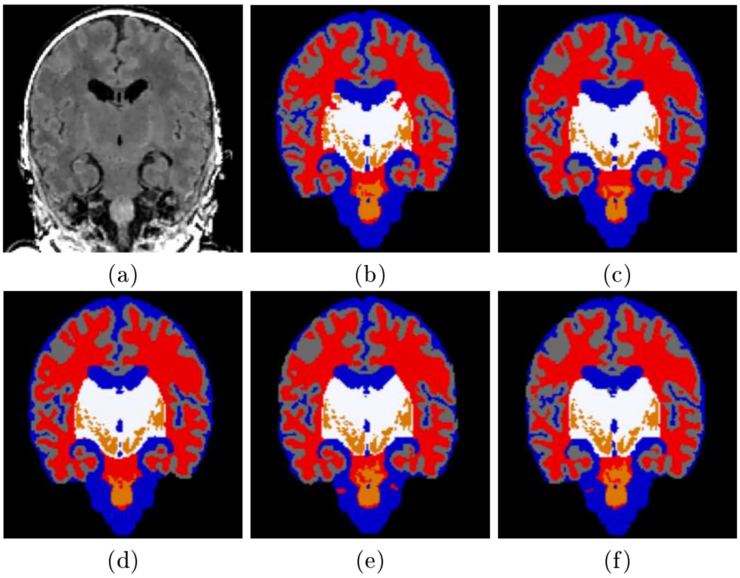
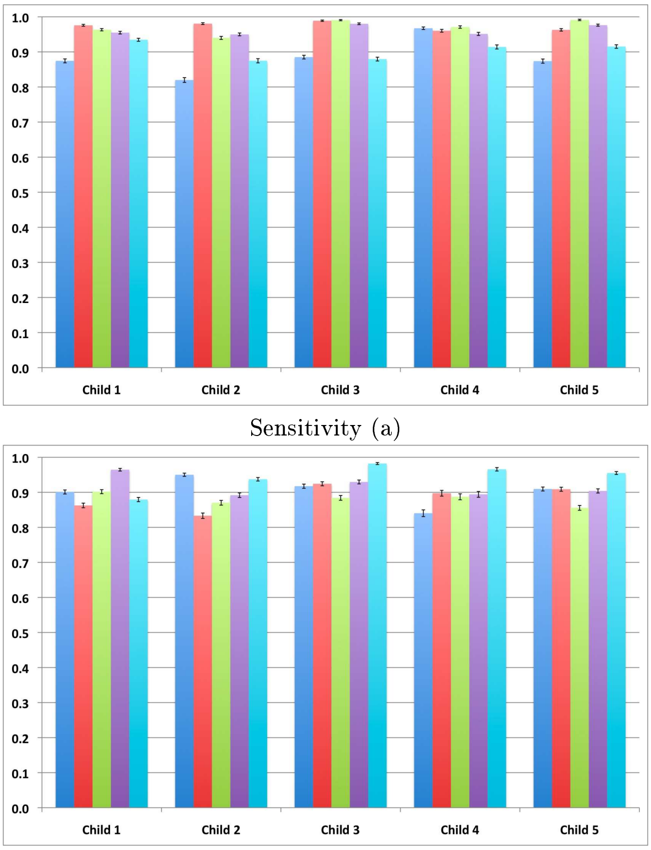


Fig. 2. Illustration of one image from the database. Coronal slice of (a) newborn T1 MRI and (b-f) its repeated manual segmentation in 5 classes done by one expert (cortical gray matter - grey, sub-cortical gray matter - white, unmyelinated white matter - red, myelinated white matter - orange - and CSF - blue).

Image Database We have then applied our algorithm to five datasets of neonate MRI segmentation (one of them illustrated in Fig. 2) selected from

MRI scans from previous studies. Each of these datasets consisted of a T1 and a T2 weighted image. After registration of the T2 image to the T1 image, five tissue classes were delineated interactively: cortical gray matter, sub-cortical gray matter, unmyelinated white matter, myelinated white matter and cerebrospinal fluid (CSF). This process was repeated five times by three experts so that for each dataset, 15 segmentations of five structures were finally available.



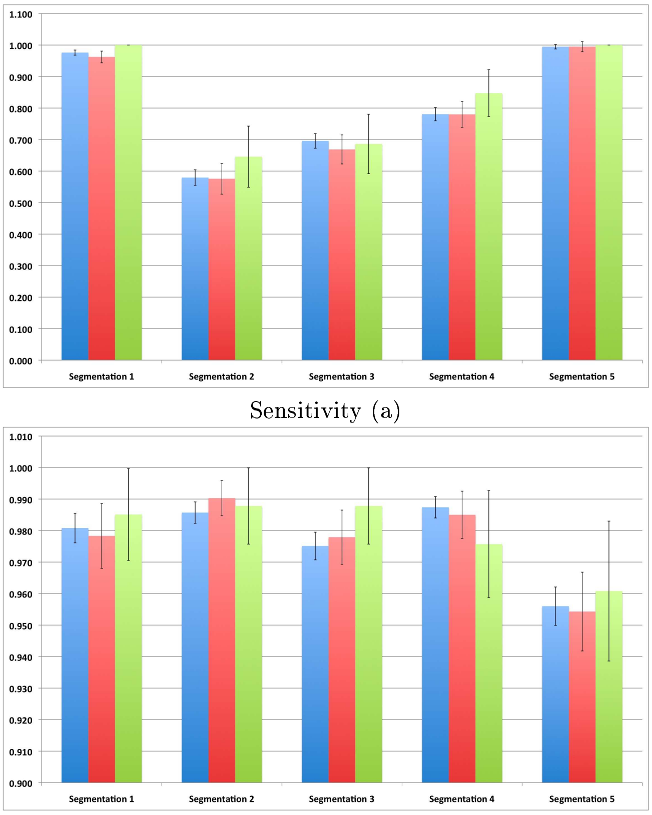
Sensitivity (a)

Specificity (b)

Fig. 3. Variability of the sensitivity and specificity parameters. Expert parameters and their variability ((a): Sensitivity, (b): Specificity) for the unmyelinated white matter segmentation. These results on five segmentations (each column of each graph) show that the standard deviations of the sensitivity and specificity parameters are low (going up to 1.3 % of the parameters values).

Evaluation of the Bounds on the Estimated Parameters We have used STAPLE for each patient on the five segmentations of one expert to determine the reference segmentation of the unmyelinated white matter for this expert, together with parameters of sensitivity and specificity for each manual segmentation. We have then used our analytical formulation to efficiently compute the observed Information Matrix for these parameters, and evaluated the covariance matrix of the parameters by simply inverting the Information Matrix.

The parameters variabilities were computed on all patients and all structures but, for clarity, we only present in Fig. 3 the results on the unmyelinated white matter, showing for each parameter its standard deviation as an error bar. This figure shows that even with only five segmentations to estimate the ground truth, the estimation of the expert parameters is very precise. The maximum relative standard deviation is indeed of 1.3 %. This however seems logical as the parameters are computed from all the voxels of the considered image.



Sensitivity (a)

Specificity (b)

Fig. 4. Influence of the image dimension on parameter variability. Standard deviation of the estimated values (bars: parameter values, error bars: standard deviations) of the sensitivity (a) and specificity (b) parameters for the image at original size (blue), subsampled once (red), and subsampled twice (green). An increase in the standard deviation values is shown as the image is subsampled.

Influence of the Image Size on Parameter Variability We also wanted to confirm in a real case previous simulated results on the influence of image size on the estimated variability of the parameters. We therefore subsampled the segmentations of one patient (again using the five segmentations of one expert) and evaluated the quality parameters as well as their variability.

We present in Fig. 4 the results of sensitivity, specificity and standard deviations (as error bars) on a patient in its original resolution, subsampled once and twice. First, we can see on some experts that the variability of their parameters becomes 0 when the images are subsampled twice. This is due to the fact that the image becomes so small that the whole region of interest for a given expert is only composed of the delineated structure, thereby removing the variability for the corresponding expert parameter. Apart from this effect, these results confirm a clear influence of the image size on the parameters bounds (error bars represent one standard deviation in Fig. 4). The standard deviations again increase when the image is subsampled. This seems quite logical as the less information is known about each expert, the more variable the parameters are.

Influence of the Number of Segmentations on Parameter Variability Finally, another potential cause of parameter variability is the number of segmentations used as an input to compute the reference segmentation. We have studied this property using binary segmentation performance estimates on ten manual segmentations of one subject. We present the evaluation of the results using from 4 segmentations up to 10 segmentations (using less experts would indeed not be meaningful for the statistical estimation of the hidden segmentation). For each number K of manual segmentations, we have performed the study over all the combinations of K images among the ten available.

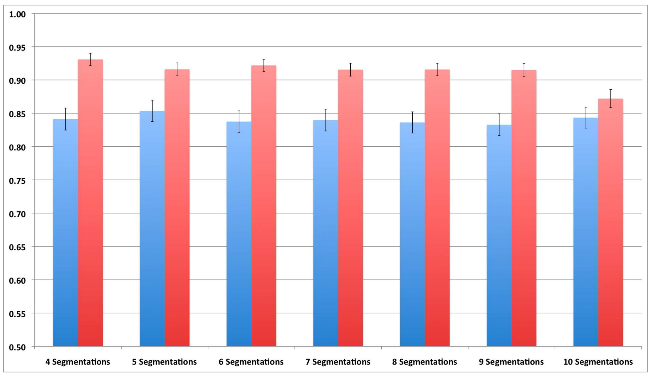


Fig. 5. Influence of the number of experts on parameter variability. Mean sensitivities (in blue) and specificities (in red) and their respective relative variability as a function of the number of experts in the study. The red bars indicate the standard deviation of the mean values over the possible combinations of K experts among the 10 available.

We present in Fig. 5 the average parameters (blue: sensitivity, red: specificity) computed over the combinations of K images. We also show (error bars on the figure) the average standard deviations for each number of experts. These results show no significant change of the variability of the parameters. This suggests that, using 4 or more experts, the size of the structure to be delineated as well as the size of the region of interest for the STAPLE computation is more influential upon the variability of the estimated parameters than the number of experts.

4 Conclusion

We have presented in this article the expression of confidence bounds on the values of the expert performance parameters computed by the STAPLE algorithm for the binary case. These formulations are based on the derivation of analytic expressions for the observed Information Matrix of the underlying EM algorithm. Such confidence bounds will be very important as they will aid in the interpretation of the performance of segmentation generators, and determine if sufficient data size and number of segmentations have been obtained to accurately characterize the performance parameters.

We have presented examples of the application of these expressions for the evaluation of confidence intervals on the estimated values of the expert parameters first in simulated experiments, showing the ability of STAPLE to obtain accurate estimates of known performance parameters. We have also utilized these expressions in the context of neonate brain segmentation, showing a dependence of the bounds with respect to the number of voxels in the region of interest for the segmentation. However, in our particular example, no correlation was detected between the number of experts and the variability of their quality parameters when using 4 or more experts in STAPLE.

This work may be further improved in the future by extending the expression of the observed Information Matrix to the multi-category case, i.e. when several structures have been segmented by each expert. This will require to take into account the interdependency between the estimated performance parameters, for example by considering only $L-1$ label independent parameters, the last one being computed from the others.

These expressions may then have many applications in terms of validation of segmentation or evaluation of intra-expert segmentation variability in a clinical context. In addition to the help to the clinician team in the assessment of the parameters determined by the STAPLE validation algorithm, this work could also be used in the future for the development of a local implementation of the STAPLE algorithm. This would allow to determine the minimal size of the region of interest required to obtain meaningful results for a given structure. Future work will then examine using this approach to evaluate spatially varying performance parameters and their bounds.

Acknowledgments

This investigation was supported in part by a research grant from CIMIT, grant RG 3478A2/2 from the NMSS, and by NIH grants R03 CA126466, R01 RR021885, R01 GM074068, R01 EB008015 and P30 HD018655.

References

1. Huttenlocher, D., Klanderman, D., Rucklige, A.: Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(9) (September 1993) 850–863
2. Chalana, V., Kim, Y.: A methodology for evaluation of boundary detection algorithms on medical images. IEEE Transactions on Medical Imaging **16**(5) (1997) 642–652
3. Dice, L.: Measures of the amount of ecologic association between species. Ecology **26**(3) (1945) 297–302
4. Jaccard, P.: The distribution of flora in the alpine zone. New Phytologist **11** (1912) 37–50
5. Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Tempany, C., Kaus, M.R., Haker, S.J., Wells, W.M., Jolesz, F.A., Kikinis, R.: Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol **11**(2) (February 2004) 178–89
6. Gerig, G., Jomier, M., Chakos, M.: Valmet: A new validation tool for assessing and improving 3D object segmentation. In: MICCAI. Volume 2208 of LNCS. (2001) 516–523
7. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7) (July 2004) 903–921
8. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. John Wiley and Sons (1997)
9. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society **39 (Series B)** (1977)
10. Meng, X., Rubin, D.: Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. Journal of the American Statistical Association **86** (1991) 899–909
11. Oakes, D.: Direct calculation of the information matrix via the EM algorithm. J. R. Statistical Society **61**(2) (1999) 479–482