Estimation of Inferential Uncertainty in Assessing Expert Segmentation Performance from STAPLE

Olivier Commowick*, Simon K. Warfield*, *Senior Member IEEE*

* Computational Radiology Laboratory, Department of Radiology,
Children's Hospital, 300 Longwood Avenue, Boston, MA, 02115, USA
E-mail: {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. For a segmentation validation task, it may be effective to compare the results of an automatic segmentation algorithm to multiple expert segmentations. Recently an Expectation Maximization (EM) algorithm for Simultaneous Truth and Performance Level Estimation (STAPLE) was developed to this end 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 the uncertainty of the estimated values. An estimate of this inferential uncertainty, if available, would allow the estimation of confidence intervals for the values of the parameters. This would facilitate the interpretation of the performance of segmentation generators, and help determine if sufficient data size and number of segmentations have been obtained to precisely characterize the performance parameters.

We present a new algorithm to estimate the inferential uncertainty of the performance parameters for binary and multi-category 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 the performance parameters are estimated by the computation of the observed Information Matrix. We use this algorithm to study the bounds on performance parameters 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 segmented data 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.

***Index Terms*—Covariance Matrix, Information Matrix, Confidence Intervals, Expectation-Maximization, Validation, STAPLE.**

I. 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 segmentations. For example, these can be the Hausdorff distance [1] or a mean distance between the two surfaces [2]. The other class of measures contains voxel-based measures, i.e. overlap measures based on voxelwise 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. Both may be used to provide insight into the quality of a segmentation [6] and to allow the comparison of segmentations. However, when validating a segmentation algorithm, using only one expert segmentation as the reference standard may be inappropriate as any individual manual segmentations have large or small errors.

In this article, we therefore focus on using several expert segmentations to estimate a reference standard and utilize it for the comparison of segmentations (from experts or algorithms). One algorithm for this, called STAPLE [7], uses an Expectation-Maximization (EM) algorithm to estimate iteratively, from a set of J expert segmentations, the hidden reference standard segmentation and performance parameters for each segmentation. These parameters characterize the agreement of a given expert with the reference standard, expressed as rates of detection of labels.

The STAPLE algorithm generates only point estimates of the performance parameters, and provides no information about the amount of uncertainty in the values of the estimates 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 precisely characterize the performance parameters. For example, consider planning to evaluate a new segmentation algorithm for a new data set or patient population. A reference standard for assessing the segmentation algorithm could be developed using repeated interactive segmentation of some images of a data set. When designing such an experiment, an estimate of the inferential uncertainty, if available, would describe confidence intervals for the values of the parameters and provide a way to determine how many experts should interactively delineate the data set and how many voxels or slices should be delineated so that the STAPLE estimates of the

parameters are precise enough (i.e. the confidence intervals are tight). Such confidence intervals indeed describe 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 as the sample variance over repeated segmentations.

We propose to estimate the covariance matrix of the performance parameters from STAPLE by computing the observed Information Matrix. This computation has been described in the general EM framework [8]. In this paper, we build upon [9] and derive, both for binary and multi-category segmentations, analytic closed form expressions necessary to compute the covariance of the performance parameters obtained from STAPLE. We then demonstrate factors influencing the uncertainty in the estimated performance parameters with simulated segmentations. Then, we apply our algorithm to characterize the segmentations of brains of newborn infants, comparing the binary and multi-category expressions, showing that our algorithm provides guidance for the design of future validation studies.

II. Method

A. The STAPLE Algorithm

We first recall briefly the principle of the STAPLE algorithm [7]. It uses as an input a set of segmentations from J experts (either manual delineations or automatic segmentations). These segmentations can either be binary segmentations or multi-category segmentations, i.e. several structures are delineated each one getting a specific label. This information is available as decisions d_{ij}, indicating the label given by each expert j for each voxel i. The goal of STAPLE is then to estimate both a reference standard segmentation $\mathbf{T}$, and parameters $\theta = \{\theta_1, \ldots, \theta_j, \ldots, \theta_J\}$ describing the agreement between the experts and the reference standard. 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 when the label of the reference standard is 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 this reference standard is unknown, an Expectation-Maximization approach [10], [8] is used to estimate $\mathbf{T}$ and the expert performance parameters. The EM algorithm proceeds by iterating 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 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 complete data log-likelihood, using the current estimate of the reference standard.

B. Covariance and Information Matrix

1) General Maximum-Likelihood Case: We are interested in the computation of confidence intervals, illustrated on Fig. 1, on the performance parameters estimated from STAPLE, i.e. a lower bound and upper bound on each estimated parameter $\hat{\theta}_{jl'l}$. This relies on the computation of the covariance matrix $\Sigma(\theta)$ of the expert parameters. This is done via the computation of the Information Matrix $I(\theta)$ of the parameters obtained after convergence of the Expectation Maximization algorithm, $I(\hat{\theta})$.

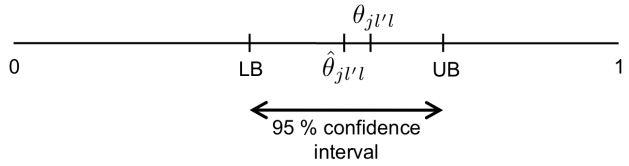


Fig. 1. **Illustration of the confidence interval on one parameter**. We aim at computing the lower (LB) and upper bound (UB) for each parameter $\hat{\theta}_{jl'l}$ estimated by STAPLE. In the case of a known ground truth (experiments on simulated data), this range can be compared to the true value $\theta_{jl'l}$ to check for the accuracy of parameter estimation in STAPLE.

If all the data was known, the computation of the Information Matrix would be simply the matrix of the second derivatives of the log-likelihood function, estimated at $\hat{\theta}$:

$$I(\hat{\theta}) = - \left[\begin{matrix} \frac{\partial^2 Q}{\partial^2 p_1} & \frac{\partial^2 Q}{\partial p_1 \partial q_1} & \cdots & \frac{\partial^2 Q}{\partial p_1 \partial q_J} \\ \frac{\partial^2 Q}{\partial q_1 \partial p_1} & \frac{\partial^2 Q}{\partial^2 q_1} & \cdots & \frac{\partial^2 Q}{\partial q_1 \partial q_J} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 Q}{\partial q_J \partial p_1} & \frac{\partial^2 Q}{\partial q_J \partial q_1} & \cdots & \frac{\partial^2 Q}{\partial^2 q_J} \end{matrix} \right] \Bigg|_{(\theta)=(\hat{\theta})} \quad (1)$$

Then, the covariance matrix is obtained using the well-known result [11] $\Sigma(\theta) = I^{-1}(\theta)$, under the assumption of a large number of samples:

$$\Sigma(\hat{\theta}) = \begin{bmatrix} \sigma^2(\hat{p}_1) & \sigma(\hat{p}_1, \hat{q}_1) & \cdots & \sigma(\hat{p}_1, \hat{q}_J) \\ \sigma(\hat{q}_1, \hat{p}_1) & \sigma^2(\hat{q}_1) & \cdots & \sigma(\hat{q}_1, \hat{q}_J) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(\hat{q}_J, \hat{p}_1) & \sigma(\hat{q}_J, \hat{q}_1) & \cdots & \sigma^2(\hat{q}_J) \end{bmatrix} \quad (2)$$

The confidence bounds of the estimated parameters are in turn computed from these values by assuming that $\hat{\theta}_{jl'l} - \theta_{jl'l}$ follows a Normal distribution $N(0, \sigma(\hat{\theta}_{jl'l}))$. A two-sided $100(1-\alpha)\%$ confidence interval can then be constructed as

$$[\hat{\theta}^{LB}_{jl'l}; \hat{\theta}^{UB}_{jl'l}] = \hat{\theta}_{jl'l} \pm z_{1-\alpha/2} \sigma(\hat{\theta}_{jl'l}) \quad (3)$$

where $z_{1-\alpha/2}$ corresponds to the z-score related to the desired confidence interval (for a 95% confidence interval, $z_{1-\alpha/2} = 1.96$). Moreover, if the Normal assumption does not hold, which may be the case when the performance parameter values are very close to 0 or 1, a function g (such as the Box-Cox transform [12] or the logit function, i.e. $\text{logit}(x) = \log(x) - \log(1-x)$) may be used to transform the parameters to obtain a Normal distribution. Assume that $g(\hat{\theta}_{jl'l}) - g(\theta_{jl'l})$ follows a Normal distribution $N(0, \sigma(g(\hat{\theta}_{jl'l})))$. Then, the confidence interval can be estimated as $[g(\hat{\theta}^{LB}_{jl'l}); g(\hat{\theta}^{UB}_{jl'l})] =$

$g(\hat{\theta}_{jl'l}) \pm z_{1-\alpha/2}\sigma(g(\hat{\theta}_{jl'l}))$. The covariance matrix computed with Eq. (2) may then be used to compute the confidence intervals: $\sigma(g(\hat{\theta}_{jl'l})) \simeq \left[\frac{\partial g}{\partial \theta}(\hat{\theta}_{jl'l})\right] \sigma(\hat{\theta}_{jl'l})$ [13] (page 626).

2) Estimating the Variance-Covariance Matrix in Missing Data Problems: In the case of an EM algorithm such as STAPLE, the hidden variables are unknown and their value may only be estimated. Therefore, 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 value of the 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. $I(\theta) = I_c(\theta) - I_m(\theta)$. These two matrices ($I_m$ and I_c) are computed from $Q(\theta|\theta^{(k)})$ once the estimates of the parameters have converged, i.e. when $\theta^{(k+1)} \approx \theta^{(k)}$. We now present the derivation of these two terms for STAPLE in the binary case, i.e. when only one structure and the background were delineated by each expert. Then, we present an extension of the observed Information Matrix computation to the multi-category case.

C. Computation of the Observed Information Matrix in the Binary Case

In the binary 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 $Q(\theta|\theta^{(k)})$:

$$Q(\theta|\theta^{(k)}) = \sum_j \sum_i \Big(W_i^{(k)} \log(\theta_{j,d_{ij},1}) + (1 - W_i^{(k)}) \log(\theta_{j,d_{ij},0}) \Big) \tag{4}$$

where $\theta_{j,d_{ij},s}$ corresponds to either θ_{j0s} or θ_{j1s} depending on the decision d_{ij} of the expert j at the voxel i. $W_i^{(k)}$ is the probability that, at iteration k, the voxel i of the reference standard $\mathbf{T}$ is labeled as 1. Using this function, we now derive the observed Information Matrix of the parameters θ.

1) 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.

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

Eq. (4) and Eq. (5) 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:

$$\mathbf{I}_{c;p_j} = \sum_i \frac{W_i^{(k)}}{\theta^2_{j,d_{ij},1}} \tag{6}$$

$$\mathbf{I}_{c;q_j} = \sum_i \frac{1 - W_i^{(k)}}{\theta^2_{j,d_{ij},0}} \tag{7}$$

2) Derivation of the Expected Missing Data Information Matrix: Once I_c has been computed, the observed Information Matrix is obtained by subtracting from I_c the expected missing data Information Matrix I_m. Computing this matrix is generally more difficult than computing the expected complete data Information Matrix. 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 [11], [8]). In the general case of any EM algorithm, an analytic expression of I_m may also be obtained by the following equation [14] if the required derivatives exist:

$$\mathbf{I}_m(\theta) = \frac{\partial^2 Q(\theta|\theta^{(k)})}{\partial\theta^{(k)}\partial\theta^T} \tag{8}$$

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:

$$\frac{\partial^2 Q}{\partial\theta_{jtt}\partial\theta^{(k)}_{nss}} = \sum_i \frac{(-1)^{1+d_{ij}}}{\theta_{j,d_{ij},t}} \frac{\partial W_i^{(k)}}{\partial\theta^{(k)}_{nss}} \tag{9}$$

where t and s are either 1 or 0, to derive the expressions for p_j and q_j. This expression gives I_m as a function of the derivatives of the probabilities of the reference standard $W_i^{(k)}$. These $W_i^{(k)}$ have been derived by Warfield et al. [7] as:

$$W_i^{(k)} = \frac{f(T_i = 1) \prod_j \theta^{(k)}_{j,d_{ij},1}}{\sum_{m=0}^{1} \left(f(T_i = m) \prod_j \theta^{(k)}_{j,d_{ij},m} \right)} \tag{10}$$

For simplicity of notation, 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 $p_n^{(k)} = \theta^{(k)}_{n11}$ and $q_n^{(k)} = \theta^{(k)}_{n00}$ can be derived:

$$\frac{\partial W_i^{(k)}}{\partial\theta^{(k)}_{nss}} = (-1)^{1+d_{in}} \pi_0\pi_1 \frac{\left(\prod_{l\neq n} \theta^{(k)}_{l,d_{il},s}\right)\left(\prod_l \theta^{(k)}_{l,d_{il},1-s}\right)}{\left(\sum_{m=0}^{1} \pi_m \prod_l \theta^{(k)}_{l,d_{il},m}\right)^2} \tag{11}$$

where s is either 0 or 1. Therefore, the expected missing data Information Matrix, defined in Eq. (8), can be computed by substituting Eq. (11) into Eq. (9). In practice, these values are

computed easily by evaluating the different expressions at each voxel.

D. An Extension to Multi-Category Labels

We now present an extension of the computation of the observed Information Matrix to the multi-category STAPLE. We therefore now consider that each expert delineates L structures labeled from 0 to $L-1$. Each expert is also associated with an $L \times L$ matrix of parameters: θ_j, as explained in Section II-A. In this case, the expected value of the complete data log-likelihood function Q is expressed as follows (see [7]):

$$Q(\theta|\theta^{(k)}) = \sum_j \sum_i \sum_s W_{si}^{(k)} \log(\theta_{j,d_{ij},s}) \tag{12}$$

As in the binary case, the performance parameters are related by the constraint that $\sum_{s'} \theta_{js's} = 1$. This constraint on the sum of the parameters on each row of the performance parameter matrix ensures that for L labels, there are only $L-1$ free variables. In the binary case, it was straightforward to select the sensitivity and specificity parameters as the variables to compute the bounds for. In the multi-category case, it is again possible to compute bounds on the $L \times (L-1)$ free parameters in each row, but this implies selecting one of the variables in each row as a fixed parameter entirely determined by the row constraint. Rather than arbitrarily select any one parameter in each row in this way, we have preferred to estimate the bounds for all $L \times L$ variables and to not utilize the row sum constraint to reduce the number of parameters.

1) Derivation of the Expected Complete Data Information Matrix: The analytical expression of $I_c(\theta)$ can be obtained from Eq. (5) and is expressed from the second derivatives of the expected value of the complete data log-likelihood Q with respect to θ. Again, only the diagonal terms are not zero due to the independence of the performance parameters and I_c is therefore composed of the terms:

$$\mathbf{I}_{c;\theta_{js's}}(\theta) = \sum_{i:d_{ij}=s'} \frac{W_{si}^{(k)}}{\theta_{js's}^2} \tag{13}$$

2) Derivation of the Expected Missing Data Information Matrix: Once the expected complete data Information Matrix is derived, we need to subtract the expected missing data Information Matrix from it to obtain the observed Information Matrix of the parameters. We derived the analytical expression of I_m from the general equation proposed in [14] for a general EM algorithm (rewritten in Eq. (8)). In the multi-category case, these second derivatives are expressed as follows:

$$\frac{\partial^2 Q}{\partial \theta_{nt't}^{(k)} \partial \theta_{js's}} = \sum_{i:d_{ij}=s'} \frac{\partial W_{si}^{(k)}}{\partial \theta_{nt't}^{(k)}} \frac{1}{\theta_{js's}} \tag{14}$$

As for the binary case, this requires to derive the expression $\frac{\partial W_{si}^{(k)}}{\partial \theta_{nt't}^{(k)}}$ for all parameters. First, we know from [7] the expression of $W_{si}^{(k)}$ as a function of $\theta^{(k)}$ parameters:

$$W_{si}^{(k)} = \frac{\pi_s \prod_j \theta_{j,d_{ij},s}^{(k)}}{\sum_m \left(\pi_m \prod_j \theta_{j,d_{ij},m}^{(k)} \right)} \tag{15}$$

where π_s correspond to the prior probability of having the structure s. From Eqs. (14) and (15), a first observation can be made on the derivatives to be computed: if $d_{in} \neq t'$, then $\frac{\partial W_{si}^{(k)}}{\partial \theta_{nt't}^{(k)}} = 0$. Otherwise, two cases arise: $t = s$ (both the numerator and denominator depend on $\theta_{nt't}^{(k)}$) and $t \neq s$ (only the denominator depends on $\theta_{nt't}^{(k)}$). These two cases lead to the following expressions. If $t = s$, the derivative is expressed as follows:

$$\frac{\partial W_{si}^{(k)}}{\partial \theta_{nt't}^{(k)}} = \frac{\pi_t \left(\prod_{l \neq n} \theta_{l,d_{il},t}^{(k)} \right) \left(\sum_{m \neq t} \pi_m \prod_l \theta_{l,d_{il},m}^{(k)} \right)}{\left(\sum_m \pi_m \prod_l \theta_{l,d_{il},m}^{(k)} \right)^2} \tag{16}$$

If $t \neq s$, then the equation changes slightly:

$$\frac{\partial W_{si}^{(k)}}{\partial \theta_{nt't}^{(k)}} = -\frac{\pi_t \pi_s \left(\prod_{l \neq n} \theta_{l,d_{il},t}^{(k)} \right) \left(\prod_l \theta_{l,d_{il},s}^{(k)} \right)}{\left(\sum_m \pi_m \prod_l \theta_{l,d_{il},m}^{(k)} \right)^2} \tag{17}$$

By substituting Eqs. (16) and (17) into Eq. (14), we are then able to compute the missing data Information Matrix and therefore the observed Information Matrix of the parameters.

3) Relationship between Multi-Category and Binary Segmentation Formulation: As mentioned above, the assumption of independence between the parameters is not true because of the constraint on some parameters to sum up to 1: $\sum_{s'} \theta_{js's} = 1$. There does not exist to our knowledge a way to take into account this interdependency in the computation of the observed Information Matrix. The two derivations presented in this paper are therefore different from each other. However, these two expressions are still related. If we consider in the binary case the full 2×2 matrix of parameters as independent, then the expression of the observed Information Matrix will be the multi-category expression. Conversely, considering the multi-category expression in the case of $L = 2$, if we consider the off-diagonal terms as exact (which can be done as they are entirely determined by the diagonal terms), then the formulation of the observed Information Matrix is exactly the binary case expression. In the multi-category case, there is no clear choice for reducing the number of free variables, and we prefer to compute the bounds for all of the variables.

III. Results

To illustrate our formulation for deriving confidence intervals of the estimated segmentation parameters, we will present two applications. First, we demonstrate the estimation of inferential uncertainty of the values of the parameters estimated from a dataset of simulated images. Then, we present the application of our framework to obtain confidence intervals for the performance parameters on a manually segmented neonate database.

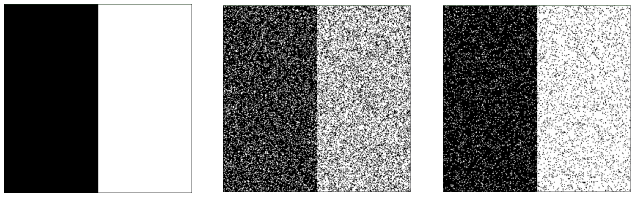
	256×256 Data		128×128 Data	
Segmentation #	Sens. (est. ; true) [LB;UB]	Spec. (est. ; true) [LB;UB]	Sens. (est. ; true) [LB;UB]	Spec. (est. ; true) [LB;UB]
1	0.7036 ; 0.7032 [0.6986 ; 0.7086]	0.8011 ; 0.8011 [0.7967 ; 0.8054]	0.6980 ; 0.6975 [0.6880 ; 0.7080]	0.7988 ; 0.7986 [0.7901 ; 0.8075]
2	0.7005 ; 0.7006 [0.6955 ; 0.7055]	0.7964 ; 0.7969 [0.7920 ; 0.8008]	0.6998 ; 0.6995 [0.6898 ; 0.7098]	0.7960 ; 0.7961 [0.7872 ; 0.8048]
3	0.7012 ; 0.7012 [0.6962 ; 0.7062]	0.7909 ; 0.7913 [0.7865 ; 0.7953]	0.6995 ; 0.6996 [0.6895 ; 0.7095]	0.7980 ; 0.7983 [0.7892 ; 0.8067]
4	0.6968 ; 0.6964 [0.6918 ; 0.7018]	0.8003 ; 0.8004 [0.7959 ; 0.8046]	0.6975 ; 0.6974 [0.6875 ; 0.7075]	0.8007 ; 0.8009 [0.7920 ; 0.8095]
5	0.7029 ; 0.7023 [0.6979 ; 0.7079]	0.8010 ; 0.8008 [0.7966 ; 0.8053]	0.6989 ; 0.6985 [0.6890 ; 0.7089]	0.7964 ; 0.7963 [0.7876 ; 0.8052]
6	0.9017 ; 0.9015 [0.8984 ; 0.9049]	0.8973 ; 0.8978 [0.8940 ; 0.9007]	0.8976 ; 0.8976 [0.8909 ; 0.9043]	0.8998 ; 0.9003 [0.8932 ; 0.9064]
7	0.9002 ; 0.8998 [0.8969 ; 0.9035]	0.8995 ; 0.8997 [0.8961 ; 0.9028]	0.8992 ; 0.8987 [0.8925 ; 0.9058]	0.9012 ; 0.9012 [0.8947 ; 0.9078]
8	0.8998 ; 0.8994 [0.8965 ; 0.9031]	0.8986 ; 0.8988 [0.8952 ; 0.9019]	0.9038 ; 0.9037 [0.8973 ; 0.9103]	0.9027 ; 0.9031 [0.8961 ; 0.9092]
9	0.8982 ; 0.8982 [0.8949 ; 0.9016]	0.9018 ; 0.9024 [0.8985 ; 0.9051]	0.8943 ; 0.8945 [0.8875 ; 0.9010]	0.8960 ; 0.8967 [0.8893 ; 0.9027]
10	0.8997 ; 0.8993 [0.8964 ; 0.9030]	0.9007 ; 0.9009 [0.8974 ; 0.9040]	0.8976 ; 0.8969 [0.8910 ; 0.9043]	0.9036 ; 0.9033 [0.8971 ; 0.9101]

TABLE I

EVALUATION OF THE EXPERT PARAMETERS CONFIDENCE INTERVALS ON SIMULATED IMAGES. EXPERIMENTS WITH SIMULATED SEGMENTATIONS SHOWING THE ESTIMATED PARAMETERS (EST.) FOR EACH SEGMENTATION, THE TRUE VALUE OF THE PARAMETER (TRUE), AND THE CONFIDENCE INTERVAL AT A 95 % LEVEL ([LB;UB]) ESTIMATED BY OUR NEW ALGORITHM. RESULTS ARE SHOWN FOR SENSITIVITY (SENS.) AND SPECIFICITY (SPEC.). THE KNOWN (TRUE) PARAMETERS FALL WITHIN THE CONFIDENCE INTERVAL OF THE ESTIMATED PARAMETERS. RESULTS ARE PRESENTED FOR 256×256 AND 128×128 IMAGES, SHOWING AN INCREASE IN THE UNCERTAINTY WHEN DECREASING THE SIZE OF THE IMAGES.

A. Experiments on Simulated Data

1) Impact of Data Size and STAPLE Precision: To evaluate our algorithm with respect to a known ground truth, we created a database of ten segmentations (2D images, size 256×256), illustrated in Fig. 2, divided into two groups. From the ground truth in Fig. 2(a), we 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 evaluate the influence of the image size on the confidence intervals of the parameters estimates, we have also generated a second database using the same parameters but with an image size of 128×128.



(a) (b) (c)

Fig. 2. **Database of Simulated Images.** Simulated images used for the validation of our confidence intervals estimation method : (a): original segmentation, (b): simulated segmentation of group 1 (sensitivity: 0.7, specificity: 0.8), (c): simulated segmentation of group 2 (sensitivity: 0.9, specificity: 0.9).

We have then run STAPLE to convergence (so that $\theta^{(k+1)} \approx \theta^{(k)}$) on the images of both databases to estimate a reference standard and utilized our algorithm to estimate the confidence intervals of the values of the parameters. The results are presented in Table I for the two databases.

Our first observation was that the non diagonal terms of the covariance matrix were always much smaller than the diagonal terms. This comes from the fact that only the expected missing data Information Matrix I_m is non-diagonal. If the reference standard was known, then the covariance matrix would be computed only as the expected complete data Information Matrix I_c, which is diagonal (see Eqs. (4) and (5)). In the STAPLE algorithm, the reference standard is not known and this leads to non zero off-diagonal terms. The figures in Table I show that all the true values of the parameters (sensitivity and specificity) fall within the 95% confidence interval around the estimated values of the parameters. Deriving the confidence bounds on these parameters therefore allows us to show that the estimation performed by STAPLE is very precise. The second observation that can be made on these figures is on the influence of the image size on the confidence bounds of the parameters. Our experiments indeed show a clear correlation between the image size and the width of the confidence interval, which is increasing when the image size is smaller.

2) Impact of the Performance Parameters Initialization: To evaluate the influence of initialization on the estimated parameters and the confidence intervals, we have used a database

			Initialization 0.9999		Initialization 0.3	
Seg #	True Sens	True Spec	Sens. (est.) [LB;UB]	Spec. (est.) [LB;UB]	Sens. (est.) [LB;UB]	Spec. (est.) [LB;UB]
1	0.2967	0.3000	0.7006 [0.6951 ; 0.7061]	0.7025 [0.6972 ; 0.7078]	0.2959 [0.2904 ; 0.3014]	0.3011 [0.2956 ; 0.3066]
2	0.3026	0.3022	0.6981 [0.6926 ; 0.7036]	0.6963 [0.6908 ; 0.7018]	0.3021 [0.2966 ; 0.3076]	0.3035 [0.2980 ; 0.3090]
3	0.3024	0.3029	0.6965 [0.6910 ; 0.7020]	0.6956 [0.6901 ; 0.7011]	0.3028 [0.2974 ; 0.3083]	0.3052 [0.2997 ; 0.3107]
4	0.2997	0.2998	0.7014 [0.6959 ; 0.7068]	0.7000 [0.6945 ; 0.7055]	0.2984 [0.2929 ; 0.3039]	0.3003 [0.2948 ; 0.3058]
5	0.3005	0.2988	0.7025 [0.6970 ; 0.7080]	0.6993 [0.6938 ; 0.7048]	0.2992 [0.2937 ; 0.3046]	0.2992 [0.2937 ; 0.3047]
6	0.2988	0.2986	0.7018 [0.6963 ; 0.7073]	0.7001 [0.6947 ; 0.7056]	0.2982 [0.2927 ; 0.3037]	0.2998 [0.2943 ; 0.3053]
7	0.3028	0.2991	0.7019 [0.6964 ; 0.7074]	0.6967 [0.6912 ; 0.7022]	0.3017 [0.2962 ; 0.3072]	0.2998 [0.2943 ; 0.3052]
8	0.2985	0.2994	0.7013 [0.6958 ; 0.7068]	0.7008 [0.6953 ; 0.7063]	0.2977 [0.2922 ; 0.3031]	0.3004 [0.2949 ; 0.3059]
9	0.2986	0.2978	0.7034 [0.6979 ; 0.7089]	0.7012 [0.6957 ; 0.7067]	0.2972 [0.2917 ; 0.3027]	0.2983 [0.2928 ; 0.3038]
10	0.7957	0.8993	0.0995 [0.0954 ; 0.1036]	0.2064 [0.2011 ; 0.2117]	0.7969 [0.7916 ; 0.8022]	0.8981 [0.8940 ; 0.9022]

TABLE II

EVALUATION OF THE INFLUENCE OF INITIALIZATION ON PERFORMANCE PARAMETERS ESTIMATES AND CONFIDENCE INTERVALS. EXPERIMENTS WITH SIMULATED DATA FOR TWO DIFFERENT INITIALIZATIONS OF THE PARAMETERS (0.3 AND 0.9999). RESULTS SHOW THE ESTIMATED PARAMETERS (EST.) FOR EACH SEGMENTATION, THE TRUE VALUE OF THE PARAMETER, AND THE CONFIDENCE INTERVAL AT A 95 % LEVEL ([LB;UB]) ESTIMATED BY OUR NEW ALGORITHM. RESULTS ARE SHOWN BOTH FOR SENSITIVITY (SENS.) AND SPECIFICITY (SPEC.). THIS TABLE SHOWS THAT THE CONFIDENCE INTERVALS DO NOT DEPEND STRONGLY ON THE INITIALIZATION BUT THE ESTIMATED VALUES OF THE PERFORMANCE PARAMETERS DO.

of ten segmentations (2D images, size 256×256) based on the same ground truth as above. In this experiment, we have generated 9 images with relatively low quality segmentations (sensitivity and specificity at 0.3) and one with good quality (sensitivity of 0.8 and specificity of 0.9). Then, we ran STAPLE on this database with two different initializations: one close to the true parameters (all estimates are initialized at 0.3) and one where we suppose all experts are good (all parameters initialized at 0.9999). We present the results of these experiments and the confidence intervals estimated in Table II.

This table clearly shows an influence of the initialization on the estimated performance parameters. When the parameters are initialized far away from their true values (0.9999), the estimated parameters converge to erroneous values for all experts. On the contrary, when the parameters are better initialized (all at 0.3), the algorithm converges to values close to the true sensitivities and specificities (which are included in the confidence intervals around the estimated performance parameters). Another very important result is that, despite this great change in the estimated values, the confidence interval widths are very similar. This demonstrates that our formulation for the computation of the confidence intervals estimates how precise the estimation of the parameters is, not the actual accuracy of these estimates.

B. Evaluation of Inferential Uncertainty of Parameters on a Neonate Database

1) Image Database: We have applied our algorithm to five datasets of neonate MRI segmentations (one of them illustrated in Fig. 3) 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 on the T1 image, five tissue classes were delineated interactively on one slice: 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 newborn MRI, 15 segmentations into five structures were available.

2) Evaluation of the Confidence Bounds of the Parameters: To evaluate intra-expert segmentation variability, we have used STAPLE for each patient on the five segmentations of one expert to determine the reference standard 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.

We computed the confidence intervals of the parameters using the binary case formulation separately on all patients and all structures. We only present in Fig. 4 a representative

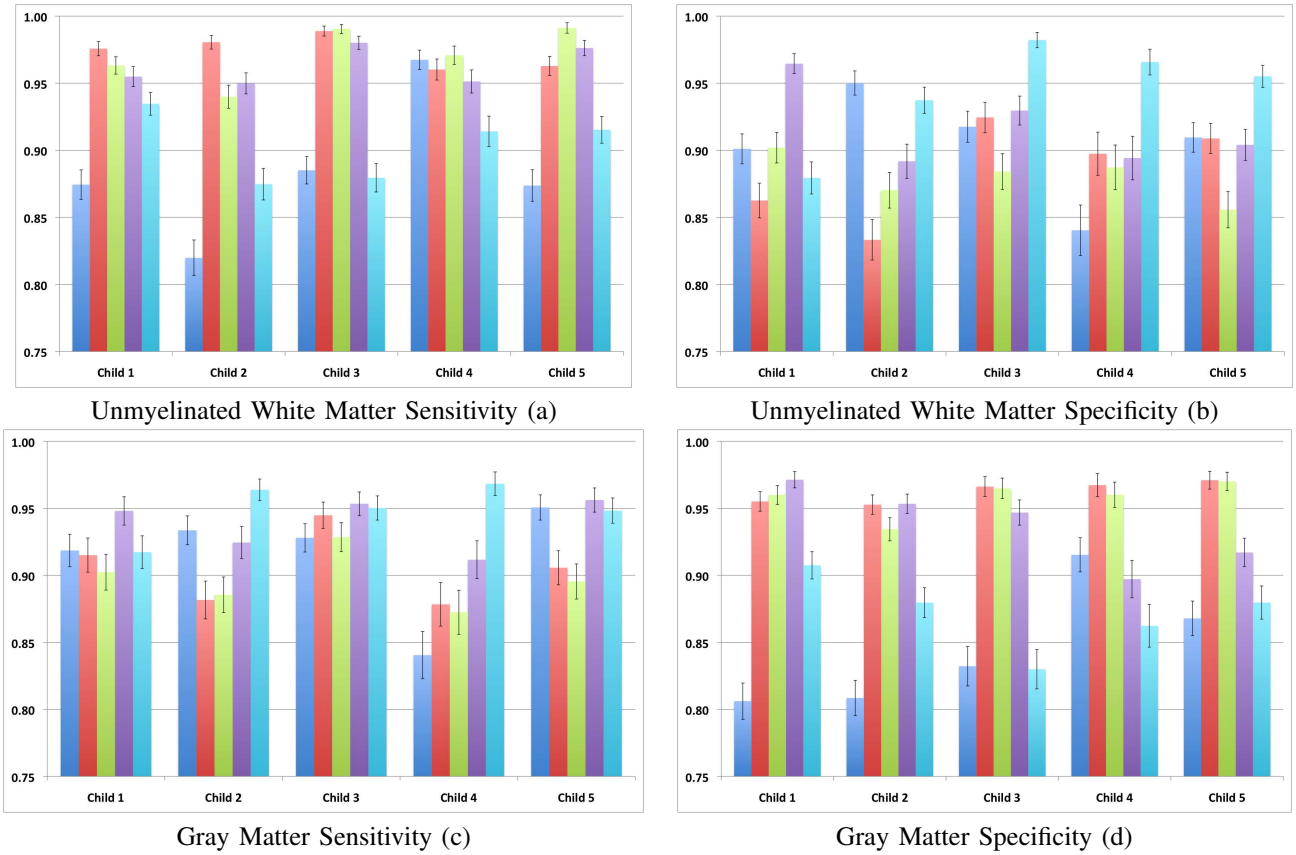


Fig. 4. **Confidence bounds of the sensitivity and specificity parameters.** Expert parameters and their confidence intervals ((a, c): Sensitivity, (b, d): Specificity) for the white matter segmentation (a, b) and the gray matter segmentation (c, d). Each child segmentations were treated separately, each column for each child represents an expert's segmentation. The results on five datasets (each column of each graph) show that the confidence intervals of the estimated sensitivities and specificities are very tight.

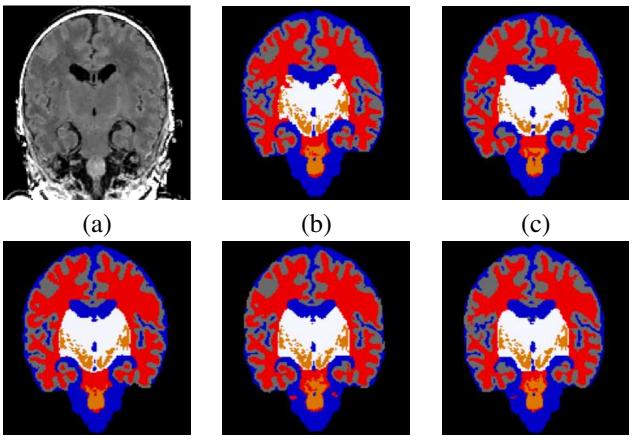
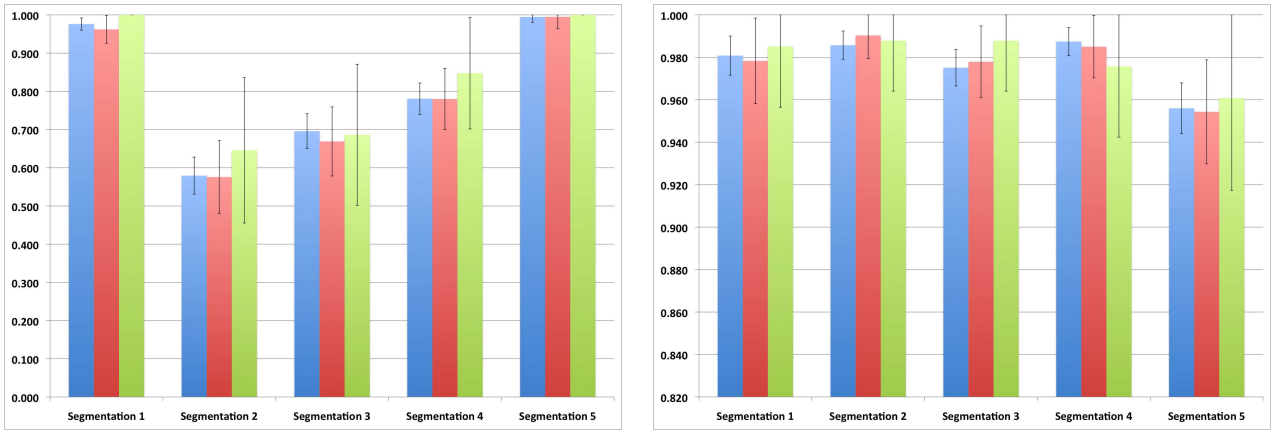


Fig. 3. **Illustration of one image from the database.** Coronal slice of (a) newborn T1 MRI and (b-f) its repeated manual segmentation in 5 classes done by one expert (cortical gray matter - grey, sub-cortical gray matter - white, unmyelinated white matter - red, myelinated white matter - orange - and CSF - blue). Other images in the database were similar to this specific example.

example of the results on the unmyelinated white matter and the gray matter for five patients using the five segmentations of one expert (each cluster in the figure illustrates independent experiments on each patient), showing for each performance parameter its confidence interval as an error bar. This figure shows that even with only five segmentations to estimate the reference standard, the estimation of the expert performance parameters is still very precise. The maximum relative standard deviation is indeed of 1.3 %.

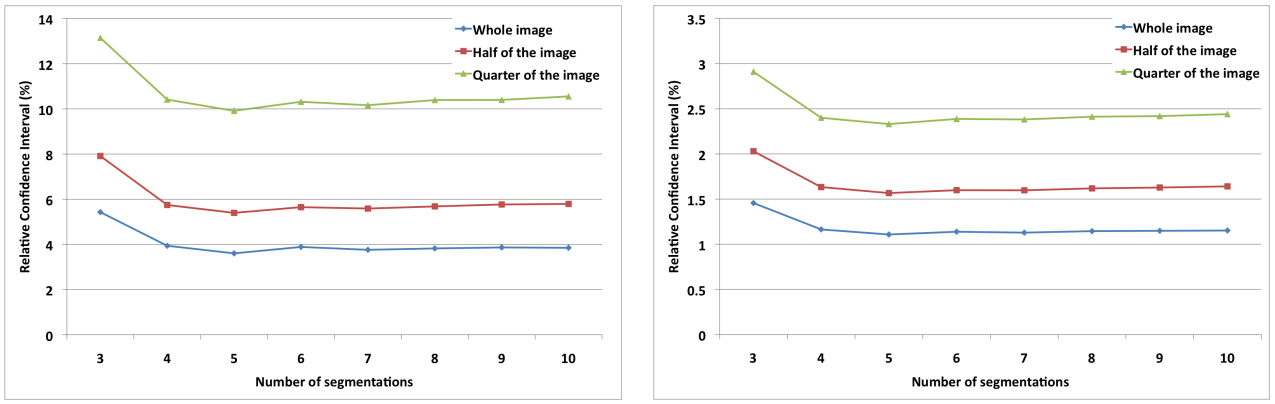
3) Influence of the Number of Voxels on the Confidence Intervals: We also wanted to confirm with data from a subject previous results on simulated data on the influence of image size on the confidence intervals of the performance parameters. To this end, we subsampled the segmentations of one patient. Because the subsampling is done using nearest neighbor interpolation, the subsampling amounts to taking one row and column every two in the image. We then ran STAPLE until convergence on the subsampled segmentations of one expert for one patient and computed the confidence intervals on the parameters. We present in Fig. 5 the results of sensitivity, specificity and confidence intervals (as error bars) on a patient in its original resolution (in blue), subsampled once (in red) and twice (in green).

First, we can see on some experts that the confidence intervals of their parameters become 0 when the image size is divided by 4 in each direction. 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.



(a) Sensitivity (b) Specificity

Fig. 5. **Influence of the image dimension on confidence bounds of the parameters.** 95 % confidence intervals on the estimated values of the sensitivity (a) and specificity (b) parameters for the image at original size (blue), subsampled once (red), and subsampled twice (green). These show a decrease in the confidence in the estimated parameters as the image is subsampled, reflecting that the confidence in the estimates decreases when the amount of available data is reduced.



(a) Sensitivity (b) Specificity

Fig. 6. **Influence of the number of experts on the confidence intervals of the performance parameters.** Average relative confidence intervals values (in percent of the average performance parameter) as a function of the number of experts used in STAPLE. For each number of experts, all combinations of K experts among the 15 available were used to compute the average. The three curves show the results using: the whole images (blue), half of the images (red), and the upper left quarter of the images (green) to compute the STAPLE performance estimates.

The standard deviations nearly double when the image is subsampled.

4) Influence of the Number of Experts on the Confidence Intervals: Another potential cause of uncertainty of the estimated values of the parameters is the number of segmentations used to compute the reference standard. We have studied this property using binary segmentation performance estimates on 15 manual segmentations of one subject. We present the evaluation of the results using from 3 segmentations up to 10 segmentations. For each number K of manual segmentations, we have performed the study over all the combinations of K images among the 15 available.

We present in Fig. 6 the average relative values (in percent of the average performance parameter) of the 95% confidence intervals for each number of experts for sensitivity and specificity parameters. These results show that the relative confidence interval decreases rapidly with the number of experts, and is stable for more than five experts. Moreover, we also present in this figure three curves, using only part of the images to estimate the performance parameters (green: a quarter of the image, red: half of the image, and blue: the whole image). 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 confidence bounds of the estimated parameters than the number of experts. Overall, both these aspects should be taken into account when designing a validation study to ensure enough experts and a sufficient region have been delineated to get precise estimates of the performance parameters for each expert.

5) Evaluation of the Multi-Category Case Algorithm: Finally, we present an application of our algorithm for the multi-category case of STAPLE. The results have been computed on all structures and all patients but for clarity, we present the results on only 5 repeated segmentations of three structures from one expert: the cortical gray matter, the sub-cortical gray matter and the unmyelinated white matter. We have then run STAPLE on these segmentations using the multi-category case implementation (using 4 classes: 3 structures plus the background). We present in Table III the results of

	BG ($\theta_{j,bg,bg}$)	CGM ($\theta_{j,cgm,cgm}$)	UWM ($\theta_{j,uwm,uwm}$)	SCGM ($\theta_{j,scgm,scgm}$)
Seg. 1 Estimate	0.9409	0.9340	0.8152	0.9243
Seg. 1 95% CI	[0.9027 ; 0.9791]	[0.8952 ; 0.9728]	[0.7852 ; 0.8452]	[0.8422 ; 1.0]
Seg. 2 Estimate	0.8666	0.8794	0.9771	0.9615
Seg. 2 95% CI	[0.8299 ; 0.9033]	[0.8420 ; 0.9168]	[0.9440 ; 1.0]	[0.8776 ; 1.0]
Seg. 3 Estimate	0.8923	0.8831	0.9341	0.9589
Seg. 3 95% CI	[0.8551 ; 0.9295]	[0.8453 ; 0.9209]	[0.9020 ; 0.9662]	[0.8752 ; 1.0]
Seg. 4 Estimate	0.9298	0.9194	0.9463	0.9280
Seg. 4 95% CI	[0.8920 ; 0.9676]	[0.8810 ; 0.9578]	[0.9138 ; 0.9788]	[0.8459 ; 1.0]
Seg. 5 Estimate	0.9212	0.9595	0.8702	0.9520
Seg. 5 95% CI	[0.8834 ; 0.9590]	[0.9199 ; 0.9991]	[0.8392 ; 0.9012]	[0.8687 ; 1.0]

TABLE III

EVALUATION OF THE MULTI-CATEGORY CONFIDENCE INTERVALS ALGORITHM. ESTIMATED PERFORMANCE PARAMETERS VALUES AND THEIR CONFIDENCE INTERVALS (CI) OBTAINED USING OUR MULTI-CATEGORY ALGORITHM ON FIVE SEGMENTATIONS FROM ONE RATER. STUDIED STRUCTURES ARE: BG: BACKGROUND, CGM: CORTICAL GRAY MATTER, UWM: UNMYELINATED WHITE MATTER, SCGM: SUB-CORTICAL GRAY MATTER.

our algorithm, showing only the estimated values and 95% confidence intervals of the diagonal parameters, i.e. the θ_{jss}, as showing the results for all parameters would produce a very large table.

The multi-category bounds estimate enables us to determine the precision of a rater performance estimate. In this precise example, the relative standard deviations of the expert performance parameters are very tight, varying between 1.7 % and 4.5 % of the respective estimated parameters values, showing that the values estimated by STAPLE are also precise in the multi-category case. The estimation of the confidence intervals of the multi-category performance parameters will allow the determination of the minimal image size and the number of experts necessary to achieve a chosen level of precision in segmentation performance assessment.

IV. CONCLUSION

We have presented in this article the expression of confidence intervals of the expert performance parameters obtained using the STAPLE validation method, both in the binary and the multiple category case. These formulations are based on the derivation of analytic expressions for the observed Information Matrix of the underlying Expectation-Maximization algorithm. Such confidence bounds will be very important for future studies as they will aid in the interpretation of the performance of segmentation generators, and in determining the minimal size and number of segmentations to precisely characterize the performance parameters.

We have presented examples of the application of these expressions for the evaluation of the inferential uncertainty of the expert parameters in experiments on simulated images, showing that the true values of the expert performance parameters fall within the confidence intervals of the estimated values of the parameters. We have also utilized these expressions in the context of neonate brain segmentation, showing a dependence of the confidence intervals with respect to the number of voxels in the region of interest for the segmentation. Moreover, we have also shown that the number of experts used in the study may influence the uncertainty of the estimated parameters. In our particular case, we have shown that, independently of the size of the segmentation, the uncertainty of the parameters is stable when 5 or more experts are used in the study. These experiments provide an important insight on the design of future experiments for segmentation validation. It will indeed be very important to have as many experts as possible when comparing small segmentations, in order to minimize the potentially large uncertainty on the values of the estimated parameters. Otherwise, if the structure of interest is large enough, using a small number of experts will not affect the inferential uncertainty in the values of the performance parameters.

Finally, we have presented experiments illustrating the multi-category formulation of the confidence bounds computation. These confidence bounds are useful for the design of future segmentation comparison experiments.

These expressions may then have many other applications in terms of validation of segmentation or evaluation of intra-expert segmentation variability in a clinical context. In addition to providing guidance in the interpretation of the parameters determined by the STAPLE validation algorithm, this work could be used in the future for the development of a spatially localized STAPLE algorithm by computing performance parameters estimates in a blockwise manner. The bounds estimated with the algorithm described here would allow us to determine the minimal size of the region of interest required to obtain precise parameter estimates 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, grants RG 3478A2/2 and RG 4032A1/1 from the NMSS, and by NIH grants R03 EB008680, R01 RR021885, R01 GM074068, R01 EB008015 and P30 HD018655.

REFERENCES

[1] D. Huttenlocher, D. Klanderman, and A. Rucklige, "Comparing images using the Hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[2] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Transactions on Medical Imaging*, vol. 16, no. 5, pp. 642–652, 1997.
[3] L. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
[4] P. Jaccard, "The distribution of flora in the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, 1912.
[5] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. C. Tempany, C. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Acad Radiol*, vol. 11, no. 2, pp. 178–89, Feb. 2004.
[6] G. Gerig, M. Jomier, and M. Chakos, "VALMET: A new validation tool for assessing and improving 3D object segmentation," in *MICCAI*, ser. LNCS, vol. 2208, 2001, pp. 516–523.
[7] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, July 2004.
[8] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley and Sons, 1997.
[9] O. Commowick and S. K. Warfield, "Estimation of inferential uncertainty in assessing expert segmentation performance from STAPLE," in *Proceedings of the 21st International Conference on Information Processing in Medical Imaging*, ser. LNCS, vol. 5636, 2009, pp. 701–712.
[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39 (Series B), 1977.
[11] X. Meng and D. Rubin, "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm," *Journal of the American Statistical Association*, vol. 86, pp. 899–909, 1991.
[12] R. Nishii, *Encyclopaedia of Mathematics*. Kluwer Academic Publishers, 2001, ch. Box-Cox Transformation.
[13] W. Meeker and L. Escobar, *Statistical Methods for Reliability Data*. John Wiley & Sons, 1998.
[14] D. Oakes, "Direct calculation of the information matrix via the EM algorithm," *J. R. Statistical Society*, vol. 61, no. 2, pp. 479–482, 1999.