Self-assessed performance improves statistical fusion of image labels

Frederick W. Bryan,^{a)} Zhoubing Xu, Andrew J. Asman, and Wade M. Allen
Electrical Engineering, Vanderbilt University, Nashville, Tennessee 37235

Daniel S. Reich
Translational Neuroradiology Unit, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892

Bennett A. Landman
Electrical Engineering, Vanderbilt University, Nashville, Tennessee 37235; Biomedical Engineering, Vanderbilt University, Nashville, Tennessee 37235; and Radiology and Radiological Sciences, Vanderbilt University, Nashville, Tennessee 37235

(Received 19 March 2013; revised 2 December 2013; accepted for publication 16 January 2014; published 14 February 2014)

Purpose: Expert manual labeling is the gold standard for image segmentation, but this process is difficult, time-consuming, and prone to inter-individual differences. While fully automated methods have successfully targeted many anatomies, automated methods have not yet been developed for numerous essential structures (e.g., the internal structure of the spinal cord as seen on magnetic resonance imaging). Collaborative labeling is a new paradigm that offers a robust alternative that may realize both the throughput of automation and the guidance of experts. Yet, distributing manual labeling expertise across individuals and sites introduces potential human factors concerns (e.g., training, software usability) and statistical considerations (e.g., fusion of information, assessment of confidence, bias) that must be further explored. During the labeling process, it is simple to ask raters to self-assess the confidence of their labels, but this is rarely done and has not been previously quantitatively studied. Herein, the authors explore the utility of self-assessment in relation to automated assessment of rater performance in the context of statistical fusion.

Methods: The authors conducted a study of 66 volumes manually labeled by 75 minimally trained human raters recruited from the university undergraduate population. Raters were given 15 min of training during which they were shown examples of correct segmentation, and the online segmentation tool was demonstrated. The volumes were labeled 2D slice-wise, and the slices were unordered. A self-assessed quality metric was produced by raters for each slice by marking a confidence bar superimposed on the slice. Volumes produced by both voting and statistical fusion algorithms were compared against a set of expert segmentations of the same volumes.

Results: Labels for 8825 distinct slices were obtained. Simple majority voting resulted in statistically poorer performance than voting weighted by self-assessed performance. Statistical fusion resulted in statistically indistinguishable performance from self-assessed weighted voting. The authors developed a new theoretical basis for using self-assessed performance in the framework of statistical fusion and demonstrated that the combined sources of information (both statistical assessment and self-assessment) yielded statistically significant improvement over the methods considered separately.

Conclusions: The authors present the first systematic characterization of self-assessed performance in manual labeling. The authors demonstrate that self-assessment and statistical fusion yield similar, but complementary, benefits for label fusion. Finally, the authors present a new theoretical basis for combining self-assessments with statistical label fusion. © 2014 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4864236>]

Key words: STAPLE, self-assessment, confidence, WebMill, rater model

1. INTRODUCTION

Expert manual segmentation is the gold standard for medical image segmentation, but such labeling of 3D imaging data is time consuming, expensive, and remains fraught with high intra- and inter-rater variability.^{1,2} Automated segmentation methods have been extraordinarily successful at consistently and efficiently finding anatomical structures in large datasets by exploiting information from intensities,³ shape models,^{4,5} atlases,⁶ or any combination of these.^{3,7,8} Yet, these meth-

ods still must be trained and validated for each new anatomy, pathology, or task, and the difficulties of manual labeling continue to plague modern research.⁹

The WebMill paradigm offers a compromise between the extensive human resources required for “expert” manual labeling and the investment in algorithm development for high-throughput automation.^{10–12} Briefly, WebMill uses large collections of nonexpert raters to collaboratively label 3D datasets and uses statistical fusion to combine the results. The idea of combining (“boosting”) a set of “weak

learners” to form a “strong learner” dates from the late 1980s¹³ and has been common since AdaBoost offered a practical, general purpose solution.¹⁴ Human-powered distributed computing continues to be applied successfully in a broad set of applications, such as the reCAPTCHA process for book digitization,¹⁵ assigning context to Twitter feeds,¹⁶ and the Amazon Mechanical Turk.^{17–19} To date, the viability of the WebMill protocol has been validated in cerebellar labeling^{10,11} and brain tumor labeling.¹² Although these efforts have made extensive use of statistical methods to categorize performance, a simple and ubiquitous avenue to assess confidence has, hitherto, not been explored—i.e., to ask the raters how confident they feel in their answers.

Herein, we present the first systematic characterization of *self-assessed performance* in manual labeling and explore use of this information in the WebMill context. We demonstrate that self-assessment and statistical fusion yield similar but complementary benefits for label fusion and present a new theoretical basis for combining self-assessments with statistical label fusion.

This paper is organized as follows. First, we develop theory to support the introduction of the new self-assessment metric into existing fusion theory. Then, we present the methods used to collect the individual observations, including an overview of the WebMill protocol. Finally, we present an experimental study of spinal cord labeling and discuss other applications and future directions.

2. THEORY

In the original WebMill experiments, observations produced by weak learners (referred to as “minimally trained raters”) were combined with STAPLER,²⁰ a robust extension of the Simultaneous Truth and Performance Level Estimation (STAPLE) framework.²¹ In order to integrate our new confidence metric into the STAPLE statistical framework, we derive a new statistical variant, Confidence Weighted STAPLER (CW-STAPLER), which enables integration of exogenous confidence information on a per-observation basis. This approach offers a different avenue for including exogenous information than other recent statistical fusion advancements.

MAP-STAPLE,²² the method most similar to ours, allows the inclusion of exogenous data as a prior in a Bayesian framework, which differs in motivation, theory, and implementation (e.g., only the estimation of the performance-level parameters is affected by the prior). Other notable variations upon the STAPLE framework include techniques that enable (1) direct estimation of task difficulty,²³ (2) estimation of spatially varying rater performance,^{24,25} and (3) nonlocal generalizations to implicitly account for spatially varying difficulty and correspondence.²⁶ With the complementary value of STAPLE-assessed and self-assessed performance established in the following experiments, it is appropriate and interesting to consider optimal learning of self-assessed behaviors and use of such information within a Bayesian context, perhaps with *Maximum a Posteriori* estimation of θ , similar to the existing MAP approaches.^{22,25} This remains an open problem and an area of interesting exploration.

2.A. Problem definition

Consider an image of N voxels, arranged as a column vector, with unknown true labels, T , where $T_i \in \{0, 1, \dots, L-1\}$, and where L is the number of labels. R “raters” (in common fusion terminology) each provide an observed segmentation of some or all of the N voxels one or more times. Let D be an $N \times R \times M$ matrix that indicates these label decisions at all N voxels. D_{ijm} indicates the decision at voxel i by rater j (one of the R raters) on the m th observation of that voxel. Each element $D_{ijm} \in \{\emptyset, 0, 1, \dots, L-1\}$ can be $\emptyset$ if that voxel was not observed by rater j for the m th time. (D is typically sparse and, in implementation, is never formed. “Unobserved” voxels that would have the $\emptyset$ value are simply skipped.)

Let $f(D|T)$ be a *generative model* that gives the probabilities of observations, given the truth. To characterize rater performance in STAPLE, let θ be an $R \times L \times L$ matrix with element $\theta_{j's's} \in [0, 1]$ that indicate the probability that rater j observes label s' at a particular voxel, given that the true label is s . The generative model now includes θ : $f(D|T, \theta)$. Let α be a new parameter with elements $\alpha_{ij} \in [0, 1]$ that indicates rater j 's self-assessed quality or “confidence” in their segmentation of voxel i . In CW-STAPLER, α is incorporated such into the generative model: now $f(D|T, \theta, \alpha)$.

2.B. Existing statistical fusion

Label fusion has been extensively covered elsewhere, including voting-based methods²⁷ and statistical fusion methods.²¹ Briefly, in majority vote²⁸ (MV), the label probabilities are determined by the number of raters who chose label s' normalized by the number of raters for each voxel (i.e., the modal choice). The mechanics of confidence weighted vote (CWV) are the same as MV, except that the raters' votes

FIG. 1. WebMill collaborative labeling tool. The web application was accessible online and available at any time for the raters' convenience. This screenshot shows a typical spinal-cord segmentation task, where the spinal cord (A) and “Confidence bar” (B) are highlighted. Note the “sad-face” (C) that was used to indicate extremely poor image quality. On the right are three actual rater segmentations that illustrate the range of rater behavior. For these examples the images have been cropped to show only the relevant portion of the image and the confidence bar corresponding to that image. The black arrows have been added in post processing to guide the reader. Spinal cord colors: brown, cyan, yellow, and orange are for the anterior, posterior, right and left white matter, and blue is for grey matter horn.

are scaled by a “confidence” metric. Herein, numerical confidence was a linear scaling from the height of the centroid of a rater’s mark on the confidence bar (Fig. 1). If the centroid of the rater’s mark was within a “sad-face” region or if the user did not mark any confidence, a trivial near-zero confidence (0.001) was assumed (nonzero to prevent their total exclusion in cases of all-sad-face).

The STAPLE framework follows the Expectation-Maximization (EM) (Ref. 29) framework to iteratively estimate both the true category labels, $\mathbf{T}$, and performance level parameters, θ .

STAPLE E-Step: Let $W_{si}^{(k)}$ be defined as the k th iteration of the conditional probability of the true label s at a voxel i , given the observations, $\mathbf{D}$, and the current estimate of the performance-level parameters, θ :

$$W_{si}^{(k)} \stackrel{\text{def}}{=} f(T_i = s | \mathbf{D}_i, \theta^{(k)}) = \frac{f(T_i = s) \prod_j f(D_{ij} | T_i = s, \theta_j^{(k)})}{\sum_n [f(T_i = n) \prod_j f(D_{ij} | T_i = n, \theta_j^{(k)})]}, \quad (1)$$

where $f(T_i = s)$ is the prior probability of label s , and $f(D_{ij} | T_i, \theta_j^{(k)})$ is the generative model. For STAPLE and STAPLER, the generative model, $f(D_{ij} = s' | T_i = s, \theta_j^{(k)}) = \theta_{js's}$, thus:

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \theta_{js's}^{(k)}}{\sum_n [f(T_i = n) \prod_j \theta_{js'n}^{(k)}]}. \quad (2)$$

STAPLE M-step: In the multilabel case, as is the case in this paper, θ_j (i.e., the performance parameters for rater j) is maximized under the constraint that $\sum_{s'} \theta_{js's} = 1$ by

$$\theta_{js's}^{(k+1)} = \frac{\sum_{i: D_{ij}=s'} W_{si}^{(k)}}{\sum_i W_{si}^{(k)}}. \quad (3)$$

STAPLER E-STEP: STAPLER, through the modification and inclusion of an extra, the product term excludes missing data and iterates over multiple observations of the same voxel:

$$W_{si}^{(k)} \stackrel{\text{def}}{=} \frac{f(T_i = s) \prod_m \prod_{j: D_{ijm} \neq \emptyset} \theta_{js's}^{(k)}}{\sum_n [f(T_i = n) \prod_m \prod_{j: D_{ijm} \neq \emptyset} \theta_{js'n}^{(k)}]}. \quad (4)$$

STAPLER M-STEP: The M-step is changed similarly, adding a second summation to be iterated over voxels that are observed more than once:

$$\theta_{js's}^{(k+1)} = \frac{\sum_{i,m: D_{ijm}=s'} W_{si}^{(k)}}{\sum_{i,m: D_{ijm} \neq \emptyset} W_{si}^{(k)}}. \quad (5)$$

2.C. Confidence weighted STAPLER (CW-STAPLER)

With the additional confidence information, α , we propose to reformulate the STAPLE generative model as

$$f(D_{ij} = s' | T_i, \theta_j^{(k)}, \alpha) = \alpha_{ij} \theta_{js's} + (1 - \alpha_{ij}) \frac{1}{L}, \quad (6)$$

where L is the total number of labels. With probability $1 - \alpha_{ij}$, all labels are chosen with equal likelihood or, with probability

α_{ij} , a label is chosen from the STAPLE model. Effectively, when a rater is “fully confident,” θ fully models the rater confusion. Let $f^{(\theta)}$ be the traditional model and $f^{(alt)}$ be the new generative model presented.

CW-STAPLER E-STEP: For ease of understanding and comparison, CW-STAPLER is developed from STAPLE, rather than STAPLER. The simple extension to CW-STAPLER is not shown. The new generative model is substituted straightforwardly into Eq. (1):

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j (\alpha_{ij} \theta_{js's}^{(k)} + (1 - \alpha_{ij}) \frac{1}{L})}{\sum_n [f(T_i = n) \prod_j (\alpha_{ij} \theta_{js'n}^{(k)} + (1 - \alpha_{ij}) \frac{1}{L})]}. \quad (7)$$

This modification causes the raters to be weighted based upon their reported confidence. Note that, in the case of perfect confidence (i.e., $\alpha_{ij} = 1 \forall i, j$), this simplifies to the original STAPLE. In cases of limiting zero-confidence, each element of W_i is multiplied by the same factor, so that that observation would have no net effect on the relative label weights within W_i , and the extreme case, the situation so that the label probabilities are entirely governed by the prior distribution, $f(T_i = s)$.

CW-STAPLER M-STEP: Maximizing θ within the generative model in the conventional way (by taking the partial derivative of the log-likelihood function) is difficult because θ cannot be separated from the sum over all voxels. An exact solution for θ would be preferred; however, here, we motivate and derive an approximate solution:

The log-likelihood function that is maximized for classic STAPLE’s $\theta_{js's}$ implicitly assumes *equal sampling* from the possible sample space. So, it is appropriate that every observation contribute equally in the maximum likelihood estimate. However, it is possible that different samples drawn at different times were drawn from different un-equal distributions (e.g., a period of poorer work or other such human impreciseness). Generative models try to model these spaces, and, thus, if the wrong one is chosen (i.e., the simple STAPLE model as opposed to the proposed CW-STAPLER model), then the estimate would be biased. To correct for this bias, we take an approach akin to importance sampling: rescaling the observations by their likelihood of occurring in one model vs the other, assuming both distributions have the same support.

Let Z_i be the scaling factor at voxel i that accounts for the sampling that occurred under the $f^{(\theta)}(D_{ij} = s' | T_i = s, \theta) = \theta_{js's}$ model versus the “true,” alternate model: $f^{(alt)}(D_{ij} = s' | T_i = s, \theta, \alpha) = \alpha_{ij} \theta_{js's} + (1 - \alpha_{ij}) \frac{1}{L}$,

$$Z_{is'} = \frac{P^{(\theta)}(s')}{P_i^{(alt)}(s')} = \frac{\frac{1}{K} \sum_m (\sum_s [\theta_{jD_{ijs}} \Pi(s)])}{\frac{1}{K} \sum_i (\sum_m [(\alpha_{ij} \theta_{jD_{ijs}} + (1 - \alpha_{ij}) \frac{1}{L}) \Pi(s)])}, \quad (8)$$

where $\Pi(s)$ is the prior on s and K is the total number of observations of that voxel. Note that the innermost product, $\theta_{jD_{ijs}} \Pi(s) = P(s' | s) P(s) = P(s', s)$, which is

marginalized over s becomes: $\sum_s \theta_{js's} \Pi(s) = \sum_s P(s', s) = P(s')$. This is then marginalized over all observations of that voxel, $1, \dots, M$, in the full expression. (Notation: $\theta_{jD_{ij}s}$ is used to indicate that the element of θ used in the summation, $\theta_{js's}$, varies by voxel.) Making a further approximation, we simplify this equation, eliminating the outer loop over m , by taking the average θ . Let:

$$\bar{\theta}_{s's} = \frac{1}{R} \sum_{j=1}^R \theta_{js's}. \quad (9)$$

The numerator of Eq. (8) does not depend on α_i and can be precomputed at the beginning of the M-step:

$$\xi_{s'} = \sum_{s=1}^L [\bar{\theta}_{s's} \Pi(s)], \quad (10)$$

$$\begin{aligned} Z_{is'} &\approx \frac{\sum_s [\bar{\theta}_{s's} \Pi(s)]}{\sum_s [(\alpha_i \bar{\theta}_{s's} + (1 - \alpha_i) \frac{1}{L}) \Pi(s)]} \\ &= \frac{\xi_{s'}}{\sum_s [(\alpha_i \bar{\theta}_{s's} + (1 - \alpha_i) \frac{1}{L}) \Pi(s)]}. \end{aligned} \quad (11)$$

The denominator is greatly simplified by distributing the sum:

$$Z_{is'} = \frac{\xi_{s'}}{\alpha_i \xi_{s'} + (1 - \alpha_i) \frac{1}{L}}. \quad (12)$$

In the original STAPLE, θ_j is found by solving:

$$\begin{aligned} \theta_j^{(k+1)} &= \operatorname{argmax}_{\theta_j} \sum_i E[\ln f(D_{ij}|T_i, \theta_j) | \mathbf{D}, \theta^{(k)}] \\ &= \operatorname{argmax}_{\theta_j} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_s E[\ln f(D_{ij}|T_i, \theta_j) | \mathbf{D}, \theta^{(k)}]. \end{aligned} \quad (13)$$

The scaling factor scales the effect of each voxel on θ_j so that

$$\begin{aligned} \theta_j^{(k+1)} &= \operatorname{argmax}_{\theta_j} \sum_{s'} \sum_{i: D_{ij}=s'} Z_{is'} \\ &\quad \times \sum_s E[\ln f(D_{ij}|T_i, \theta_j) | \mathbf{D}, \theta^{(k)}]. \end{aligned} \quad (14)$$

Ultimately:

$$\theta_{js's}^{(k+1)} = \frac{\sum_{i: D_{ij}=s'} (Z_{is'} W_{si}^{(k)})}{\sum_i (Z_{is'} W_{si}^{(k)})}. \quad (15)$$

Note the intuitive appeal of this formulation: The numerator is a scaled version of W_{si} that depends on confidence metric, α_i , and the denominator functions to normalize $\theta_{js's}$.

3. INITIALIZATION AND IMPLEMENTATION DETAILS

In all STAPLE variants, consensus voxels (i.e., voxels where all raters agree) were ignored^{23,26} and $\mathbf{W}$ was initialized using a MV. The prior, $f(T_i = s)$, was set to the relative proportions of each label present in the nonconsensus regions. Convergence was detected by measuring the difference of the average trace of the θ_j matrices between iterations. In CW-STAPLER, a set of θ_j 's is needed before $\bar{\theta}_{s's}$

can be calculated. To achieve this, the STAPLER algorithm was allowed to run for one iteration. All algorithms described above were implemented in MATLAB (Mathworks, Natick, MA). STAPLER and CW-STAPLER execution wall-times on a Dell T3500 workstation (4 core CPU, 12 GB RAM) for the large experiment were approximately 20 min (all data held in RAM).

4. IMAGING DATA

All structural intensity data were obtained on a 3T Philips Achieva (Philips Medical Systems, Best, The Netherlands) using a single channel body coil for transmission and a 16 channel neurovascular coil for signal reception. T2*-w data were obtained using a 3D gradient echo (TR/TE/a = 121/12 ms/9°) with a multi-shot EPI (EPI factor = 3) covering a field of view = $190 \times 224 \times 90$ mm³ (anterior-posterior, right-left, and foot-head) with nominal resolution = $0.6 \times 0.6 \times 3$ mm³ (30 slices/volume). Fat saturation was implemented by using a 1331 binomial excitation (ProSet). Two signal averages and a SENSE factor of 2 were used. The T2*-w acquisition was a part of a magnetization transfer (MT) experiment which acquired two echo trains for a single field of view, one with and one without a MT prepulse. The center of the imaging volume was aligned to the space between the 3rd and 4th cervical levels. Total scan time for the acquisition (two dynamics = MT prepulse off and on) was 7 min, 15 sec. Note, one of the primary reasons these data were chosen for a Web-Mill study is because it is very difficult to apply typical automated methods (e.g., multiatlas segmentation³⁰) and often requires application-specific approaches that are tailored to spinal cord anatomy (e.g., Ref. 31).

5. METHODS

5.A. Creation and validation of gold-standard

Before data review, gold-standard manual segmentations were created by a graduate student with experience in both manual medical image segmentation and spinal cord anatomy. The expert rater had access to both the T2*-w and MT images with the hope that with the two contrasts, he could determine the true anatomy of the difficult-to-image spinal cord. The expert rater labeled the T2*-w images. The segmentations were created using the Medical Image Processing, Analysis and Visualization (MIPAV) (Ref. 32) tool on using a Wacom (Wacom Tech. Corp, Vancouver, WA) tablet for high resolution input. The completion of the 67 volumes (includes a tutorial volume) took approximately 88 h.

We later conducted an inter-rater variability study to determine an upper bound on the WebMill raters' segmentations. A second experienced rater labeled a 60 slice subset from the existing set of gold-standard-labeled volumes. The slices drawn for this inter-rater study were drawn randomly from the available truth slices, but were evenly distributed thought the anatomy to account for spatially varying difficulty. For these 60 slices, the overall mean Dice Similarity Coefficient (DSC) (Ref. 33) was 0.66, which is indicative of the difficulty

in labeling small structures using low-contrast images. Note that the second rater and the millers did not have access to the T2*-w images, and used the MT images solely.

5.B. WebMill study

Raters for WebMill were recruited from the Vanderbilt University general undergraduate body, and were compensated proportional to time spent labeling (measured by Web-Mill in terms of open-window time and mouse-movement). Raters attended a 15 min training session (paid for 1 h to include transportation time) that included handling of paperwork, general (e.g., use of “paint” tool) and experiment-specific (e.g., confidence-bar) instruction in the use of the web-app tool, and examples of proper segmentations. Specifically, raters were instructed to do their best for each slice, and indicate how well they thought they did by marking the confidence bar with the paint tool. If they felt that the quality of the image was too poor to produce an accurate segmentation, they were instructed to mark the “sad-face.” Extra resources that the raters might have found useful, such as descriptions of the various anatomies, were provided as reference links within the WebMill web-app, but raters were not required to access them.

During the experiment, 2D slices were presented to the raters randomly, but raters were allowed to “peek” up and down one slice. This method ensured the privacy of the imaging data while still allowing the raters to access the useful-neighborhood slice images. The “confidence bar” was superimposed along the right-hand border of the image to be labeled [Fig. 1(b)]. Note the “sad-face” near the bottom of the confidence bar [Fig. 1(c)].

The database backbone ensured that slices were “covered” (i.e., viewed and labeled) evenly. Each slice was covered at least three times. This study was conducted with the results of¹² in mind, in which it was determined that after three-four independent segmentations, the marginal improvement of the fused image per-additional-segmentation is significantly diminished.

Over the course of five months, 86 independent raters labeled 8825 axial slices in 81 separate MRI volumes. Of these 86 raters, 61 rated more than 10 slices (i.e., 25 effectively quit the study). Of the 81 volumes, 67 were of high enough quality for an expert to confidently delineate the “true” spinal cord internal structure. (Note that even within these useful volumes, the “top” and “bottom” slices were often very blurred and could not be rated by the expert rater). A single expert-labeled volume was excluded from analysis so that it could be used as a tutorial dataset. Of the 2430 slices collected in the study, 1514 were expertly labeled, which allowed us to use 5312 of the 8825 rater-created label masks (Fig. 3) in this study.

Numeric confidence was calculated by determining the relative height of the centroid of the raters’ marks on the confidence bar. High marks on the bar became confidences near one, while marks near the bottom of the bar became near-zero. Centroids detected within the sad-face region were treated as a different class of observations, and are referred to as “sad-face” throughout this paper. Rater self-assessment

FIG. 2. Summary of rater observations. This figure includes all rater observations (slices) of slices for which we had an expert segmentation. From left to right are the three “classes” of reported confidence: none-reported, “sad-face,” and confidence reported on bar (non-“sad-face”). The curved black arrows pointing at the horizontal axis indicate the observations with DSC = 0. In the right-most plot, the red, dashed line is a linear regression that includes the DSC = 0 data. Note the lack of fit, compared to the solid, yellow line, which is a linear regression to only the observations that had mask content. The corresponding correlation coefficients are $\rho = 0.72$ and $\rho = 0.179$, respectively. All percentages are given in terms of the number of observations.

did not linearly scale with the empirical performance (Fig. 2). To approximately account for this mismatch, observed (raw) confidence metrics were projected onto a sigmoid function:

$$\alpha = f(a_{\text{raw}}; x, y, F_{\min}, F_{\max}) = \frac{1}{1 + e^{-x(a_{\text{raw}} - y)}} * (F_{\max} - F_{\min}) + F_{\min}, \quad (16)$$

which rescales the confidences into a sigmoid function with asymptotic range $[F_{\min}, F_{\max}]$, inflection point, y , and “sharpness” parameterized by x . For this work: $F_{\min} = 0.1$, $F_{\max} = 0.9$, $y = 0.1$, and $x = 10$. The effect of this mapping and its parameters are discussed later.

5.C. Secondary small study

After the large-scale study discussed above, we conducted a second study in order to observe the change in accuracy with increased number of coverages (i.e., the number of label observations at each voxel). In this small study, we were not able to recruit from the same cohort as before; the raters recruited had greater experience with medical imaging and anatomy, although not necessarily that of the spinal cord. Raters were given the same instructions as the previous cohort (via recorded video instruction), and data were collected through the same WebMill interface. Ten raters produced 417 masks that covered a small subset of the original data: 16 central slices from 3 volumes, 48 slices total. Each slice was covered 8-9 times. Confidence was projected onto the same sigmoid as before.

We Monte Carlo resampled the data to create ten data-subsets for level of coverage tested, (1-8 coverages, 80 datasets subsets total). Data were sampled without replacement within each data subset. Data were fused at the dataset level, as in the larger experiment. Accuracy was evaluated on

FIG. 3. Metrics of rater observations across space, inter-raters, and intra-raters. The leftmost plot demonstrates the remarkable consistency of rater segmentations across the space of a volume, as measured by slice number. The middle boxplot shows the range of intra- inter-rater quality as measured by the mean-DSC of their segmented slices. The boxplots only include observations with valid masks ($DSC > 0$), and the lower bar graph shows the number of both valid and empty segmentations. The first few raters rated less than three slices, so a box was not generated. The rightmost histograms show the intra-rater correlation between reported confidence and mean-DSC, ignoring sad-face (top) and ignoring sad-face and $DSC < 0.15$ slices (bottom). Red boxes indicate the correlation coefficients that may also be considered statistically significant ($p < 0.10$).

the slice-level rather than the volume level, giving 480 data-points per coverage level.

5.D. Study results

For all results, accuracy was assessed using the Dice Similarity Coefficient (DSC),³³ a metric of label overlap. All significance testing was performed using the nonparametric Wilcoxon Signed Rank test.³⁴

5.E. Data insights

In the large WebMill study, for the 5312 slices for which we had expert labels, (referred to as the “useful slices”), the vast majority (5217 or 98% of total) of raters reported some metric of confidence by marking the confidence bar or sad face (Fig. 2). Unfortunately, among these slices, which represented the subset of good-quality slices (that *should* have been labeled) there were many cases in which one or more of the raters decided that there was not enough information to create a mask, and marked the “sad face.” The bar graphs of Fig. 3 illustrate the ratio of empty to nonempty slices.

Of the 5217 observations with confidence marked, 4825 had confidence marked on the bar, rather than the sad face, and of these, most (3775/4825) had a nonempty mask, but a significant portion (1050/4825) were submitted with the anatomy unlabeled ($DSC = 0$) and only the confidence bar marked. These empty slices had confidences clustered around or less than 0.15, which corresponds to the height of the sad-face in the image. Of those in which the sad face was intentionally marked, almost all segmentations (387/392) were completely blank ($DSC = 0$), which justifies the separate classification of the “sad-face” slices from the rest of the results. Of the few slices (95) in which no confidence was indicated, most of the slices (75/95), appeared to have valid segmentations.

The quality of the rater’s observations was consistent across anatomy [Fig. 3(a)] and did not vary substantially with increasing numbers of slices rated [Fig. 3(b)]. Raters exhibited different variance in quality [Fig. 3(b)]. If viewed one-at-a-time, and ignoring the low-quality ($DSC < 0.15$) slices, the correlation between actual quality and confidence varied greatly by rater. In some cases, the correlation between a rater’s DSC and confidence was negative, suggesting that some raters were unable to self-rate effectively beyond a binary level of complexity [Figs. 3(c) and 3(d)].

5.F. Fusion results

For comparison, these fusion methods were evaluated.

- MV: The mode rater decision for each voxel was selected.
- CWV: Votes were weighted by the level of confidence that each rater specified on the confidence bar (from 0 to 1). The label with the highest total was selected.
- CWV, rescaled—same as CWV, except that the confidence metrics were rescaled to the same sigmoid used for CW-STAPLER (other rescalings were also tried with no more success, not shown).
- Random slice—To demonstrate the range of quality possible without collaboration and fusion, volumes were generated by randomly selecting each slice from the set of rater observations for that slice (300 Monte Carlo iterations with replacement).
- Highest confidence—Volumes were generated by selecting the highest-confidence slice for each slice in a volume.
- Ideal voting—Volumes were generated by determining the voting weights that minimized the error between the fused result and the truth. This method is the upper bound for global-weighted voting, though it can be

FIG. 4. Voting fusion results. Majority vote (MV) provides a lower bound, and does not even consistently exceed the performance possible by randomly selecting a slice. Incorporating confidence as a weight in voting (CWV) produces statistically significant improvement over MV and the other slice selection methods, except the ideal voting method, which was developed with posterior knowledge of segmentation quality and is the upper bound for voting methods.

beaten by the STAPLE methods, which can “weight” the labels independently.

- STAPLER—All volumes were generated simultaneously with the STAPLER algorithm discussed above.
- CW-STAPLER—All volumes were generated simultaneously with the CW-STAPLER algorithm discussed above.

Incorporating confidence into CWV results significantly higher ($p < 0.01$) fusion accuracy. In CWV, rescaling of the confidence does not improve final segmentation quality ($p > 0.05$). CWV performs better than the random slice selection and highest confidence selection methods, despite the use of the same information (Fig. 4). For the STAPLER fusions, all 66 volumes were fused simultaneously with one confusion matrix, θ_j per rater. CW-STAPLER was statistically significantly better than both STAPLER ($p < 0.05$) and CWV ($p < 0.05$). CW-STAPLER had the most effect in the lowest-quality volumes, seen as outliers in the box plots (Fig. 5). Lastly, qualitative results (Fig. 6) support the quantitative improvement produced via the incorporation of the self-assessment into the label fusion framework.

5.G. Coverage (small study) results

The results from our small, increased-coverage study are presented in Fig. 7. In the secondary study, the raters were much more experienced than the first study; therefore, MV was quite effective. Notably, CWV and MV were essentially identical for odd numbers of coverages, which further indicates all rater performance was effectively equivalent. For three raters, MV produced the best fusion. CW improved CW-STAPLE through reduction in variance. For four or more coverages, both CW-STAPLE and STAPLE outperformed the other methods. Variance was much reduced for all methods with four or more coverages.

6. DISCUSSION

Raters’ overall quality, as measured by their sets of rated slices, is somewhat lower than seen in previous WebMill experiments,^{10,12} which may be due to the magnification effect on error when comparing smaller anatomies (e.g., spinal cord) to larger ones (cerebellum, brain). Note that in a post-experiment inter-rater study, expert manual overlaps exhibited a mean-DSC of only 0.66; whereas expert inter-raters DSC

FIG. 5. Statistical fusion results. The STAPLER and confidence weighted STAPLER (CW-STAPLER) methods are shown superimposed on the previous MV, CWV, and ideal selection methods for comparison. STAPLER achieves similar results to CWV, despite not having the confidence information. CW-STAPLER achieves superior ($p < 0.05$) results, which can be appreciated in the reduced outliers in the boxplot.

FIG. 6. Qualitative fusion results. A representative slice shows the effect of the additional confidence parameter on the output of both the voting and statistical fusion techniques. The three rater observations (one submitted empty) are shown. In this example, CWV and “Best Pick” would result in an identical segmentation.

values of near 0.95 are typical in cerebral applications. For slices in which raters reported confidence, reported confidences and mean-DSC showed low, but significant correlation ($p < 0.01$) when a nonempty mask was specified (Fig. 2). Rater consistency across axial slices (Fig. 3) was unexpectedly uniform across slices but may be explained by the implicit filtering of unusable data in the expert labeling process. With the low correlation of all the observations to all the confidence measurements, the relatively large correlation coefficients observed in Fig. 3(c) were unexpected. However, when all low-quality (DSC < 0.15) slices are also ignored, we see the same effect as was observed in Fig. 2. The largest correlation coefficients were due to a cluster of low-quality, low-confidence observations that increased correlation. We considered the problem of different confidences meaning different things to the different raters, and examined rescaled confidences (e.g., rescaled using linear and z-score transformations). However, these did not alter the segmentation results (Fig. 4) nor the overall correlations seen in Fig. 2. As a final note, from Figs. 4 and 5, it is apparent that some volumes were generally more difficult than others.

FIG. 7. Fusion results versus number of coverages: Data were fused on the volume level with increasing number of coverages in 10 Monte Carlo trials. DSC for MV (computed and presented at the slice-level) improves asymptotically. Other fusion algorithm performance is given relative to MV performance. $\mu\Delta$ denotes the mean of the differences between each method and MV across all slices, volumes, and Monte Carlo iterations. $\sigma\Delta$ denotes the standard deviation of these same differences. Note that one coverage implies random selection, and data for two coverages are not shown because fusion is meaningless with two coverages.

The results from the secondary study correspond with those from the larger study. CW-STAPLER provides useful information that is impactful in the low-coverage case. For higher coverages, CW-STAPLER and STAPLER converge to answers of equivalent quality. This experiment also unsurprisingly shows that more coverages results in higher quality fused data. In future WebMill experiments or in other crowd-sourced paradigm methods, it is advisable to collect a greater depth of coverages. Mean-DSC in the small study was much improved compared to the large study. This effect is unsurprising considering the increased experience of the raters and the resulting homogeneity of the small dataset. Additionally, these raters rated the T2*-w contrast instead of the MT contrast that the inter-rater and large study used.

In each successive WebMill experiment, we learned how to better administer recruit, train, and fuse data from minimally trained raters. This study was the first to collect the confidence meta-data and to pseudo-randomly select tasks to ensure even coverages for all slices. In this study, we also allowed raters to “opt-out” of rating a particular slice because it was “unrateable” by marking the “sad-face.” Ultimately, we treated “sad-face” and low-confidence slices as identical, so there was little gained by the mixed-modal confidence reporting, and in retrospect, this process may have confused some raters. Alternatively, with knowledge of some characteristics of the correct segmentation, we could have filtered the results and created our own meta-data to augment the pseudo-random selection process. For example, in these data, it was guaranteed that each slice should have the five spinal-cord labels (defined in Fig. 1 caption), as there were no slices that did not have a spinal cord present. More robust mechanisms to ensure that “equal information” is acquired for each task/slice (as opposed to equal coverages) is of definite value to the collaboration process and promises to be an interesting area of continued practical and theoretical innovation.

Human “confidence” is intrinsically a qualitative trait and dependent on past experiences, visual cues, fatigue, and ego. Raters were effectively presented a graphical version of a Likert scale with familiar “happy face” and “unhappy face” logos at the top and bottom with green/yellow/red color scale to maximize similarity to real-life happiness assessments. We expected that individuals would self-calibrate such that relative happiness would be reflected consistently across observations from an individual rater, but the scales would be biased across raters (i.e., as with Netflix user ratings^{35,36}). In the context of WebMill image labeling, we did not observe a strong intra-rater correlation with actual performance beyond the extremes of performance (Fig. 3). This is an interesting and unexpected finding. Some characterization of the confidence data may be necessary in order to effectively incorporate the meta-data into the fusion framework. Alternatively, more experienced raters might have greater experience to draw from and be better suited for fine-grained performance estimation than raters drawn from the general public. In this experiment, we projected our data onto a sigmoid function, whose parameters were found empirically. (Though, the inflection point corresponded to the top of the sad-face, which could have been known *a priori*.) We recognize that this choice was arbitrary

and may not result in optimal fusion for other cases. However, the specifics of the confidence metric scaling/calibration should not detract from the main point: *self-assessed rater confidence is a different source of information that complements information that is leveraged in existing statistical fusion methods.*

As a first foray into using self-assessed performance, we made suboptimal choices. The limitations of the study design do not create straw-man situations or otherwise invalidate the performance gains, and, in fact, underrepresent the potential of self-assessed performance. The results of this study—both in terms p -value and experimental setup—can inform and improve future efforts.

7. CONCLUSION

This study has demonstrated that user-generated confidence metrics can be readily acquired and have practical value. This experiment was also unique in that it was the first time in which the STAPLER algorithm was applied to simultaneously fuse a large, multivolume dataset. The self-assessed confidence provides much of the benefit of statistical fusion to the direct voting methods (without the computational expense of EM iterations). We presented a simple but practical model in which the self-assessed confidence (or any exogenous sources of accuracy information) can be included in the STAPLE statistical fusion framework. Inclusion of both self-assessed confidence and statistical performance estimation increases the fusion accuracy. Improvements to the paradigm could include either rater self-assessment “training” or the use of catch trials (i.e., inserting known segmentations into the rated-slice stream) to quantitatively determine quality, which has been investigated previously²⁰ and could be used to develop a per-rater calibration scale. The relative costs and benefits of these approaches would need to be evaluated.

ACKNOWLEDGMENTS

The authors are grateful for the imaging sequences developed by Dr. Seth Smith and acquired by Peter Calabresi. The authors would also like to thank the WebMill and secondary study volunteers for their time and effort. This research was supported by in part by NIH 1R21NS064534, NIH 1R03EB012461, NIH 1T32EB014841, the NIH/NINDS Intramural Research Program, and the National Multiple Sclerosis Society (Peter Calabresi). The authors appreciate the insightful comments and assistance of the editors in improving this paper.

^{a)}Author to whom correspondence should be addressed. Electronic mail: frederick.w.bryan@vanderbilt.edu

¹M. Kaus *et al.*, “Segmentation of meningiomas and low grade gliomas in MRI,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI’99* (Springer Berlin Heidelberg, 1999).

²T. Kapur *et al.*, “Segmentation of brain tissue from magnetic resonance images,” *Med. Image Anal.* **1**(2), 109–127 (1996).

³M. Prastawa *et al.*, “Automatic brain tumor segmentation by subject specific modification of atlas priors,” *Acad. Radiol.* **10**(12), 1341–1348 (2003).

⁴S. Amari, A. Cichocki, and H. H. Yang, *A New Learning Algorithm for Blind Signal Separation*, Advances in Neural Information Processing Systems Vol. 8 (MIT Press, Cambridge, MA, 1996), pp. 757–763.

⁵D. L. Collins *et al.*, “Automatic 3-D model-based neuroanatomical segmentation,” *Human Brain Mapp.* **3**(3), 190–208 (1995).

⁶J. L. Tanabe *et al.*, “Tissue segmentation of the brain in Alzheimer disease,” *Am. J. Neuroradiol.* **18**(1), 115–123 (1997).

⁷J. J. Corso *et al.*, “Efficient multilevel brain tumor segmentation with integrated Bayesian model classification,” *IEEE Trans. Med. Imaging* **27**(5), 629–640 (2008).

⁸O. Tsang *et al.*, “Comparison of tissue segmentation algorithms in neuroimage analysis software tools,” in *Engineering in Medicine and Biology Society (EMBS), 30th Annual International Conference of the IEEE*, Vancouver, BC (IEEE, 2008).

⁹T. C. Huang *et al.*, “Semi-automated CT segmentation using optic flow and Fourier interpolation techniques,” *Comput. Methods Programs Biomed.* **84**(2-3), 124–134 (2006).

¹⁰A. X. Falcão *et al.*, “User-steered image segmentation paradigms: Live wire and live lane,” *Graph. Models Image Process.* **60**(4), 233–260 (1998).

¹¹B. A. Landman *et al.*, “Foibles, follies, and fusion: Web-based collaboration for medical image labeling,” *Neuroimage* **59**(1), 530–539 (2012).

¹²M. J. Cardoso *et al.*, “LoAd: A locally adaptive cortical segmentation algorithm,” *NeuroImage* **56**(3), 1386–1397 (2011).

¹³M. Kaus *et al.*, “Adaptive template moderated brain tumor segmentation in MRI,” *Bildverarbeitung fuer die Medizin* 102–106 (1999).

¹⁴M. Prastawa *et al.*, “A brain tumor segmentation framework based on outlier detection,” *Med. Image Anal.* **8**(3), 275–283 (2004).

¹⁵G. Cohen *et al.*, “Segmentation techniques for the classification of brain tissue using magnetic resonance imaging,” *Psychiatry Res.: Neuroimaging* **45**(1), 33–51 (1992).

¹⁶D. L. Pham, and J. L. Prince, “Adaptive fuzzy segmentation of magnetic resonance images,” *IEEE Trans. Med. Imaging* **18**(9), 737–752 (1999).

¹⁷R. M. Gardner, D. L. Brown, and R. Boice, “Using Amazon’s Mechanical Turk website to measure accuracy of body size estimation and body dissatisfaction,” *Body Image* **9**(4), 532–534 (2012).

¹⁸W. Mason and S. Suri, “Conducting behavioral research on Amazon’s Mechanical Turk,” *Behav. Res. Methods* **44**(1), 1–23 (2012).

¹⁹D. G. Rand, “The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments,” *J. Theor. Biol.* **299**, 172–9 (2012).

²⁰B. A. Landman *et al.*, “Robust statistical fusion of image labels,” *IEEE Trans. Med. Imaging* **31**(2), 512–522 (2011).

²¹S. 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 Trans. Med. Imaging* **23**(7), 903–921 (2004).

²²J. Talairach, *Atlas d’anatomie stéréotaxique du télencéphale: études anatomo-radiologiques [Atlas of Stereotaxic Anatomy of the Telencephalon]* (Masson et Cie (Masson & Co.), 1967).

²³A. Asman and B. Landman, “Robust statistical label fusion through consensus level, labeler accuracy and truth estimation (COLLATE),” *IEEE Trans. Med. Imaging* **30**(10), 1779–1794 (2011).

²⁴A. J. Asman and B. A. Landman, “Formulating spatially varying performance in the statistical fusion framework,” *IEEE Trans. Med. Imaging* **31**(6), 1326–1336 (2012).

²⁵A. J. Asman *et al.*, “Foibles, follies, and fusion: Assessment of statistical label fusion techniques for web-based collaborations using minimal training,” *Proc. SPIE* **7962**, 79623G (2011).

²⁶O. Commowick, A. Akhondi-Asl, and S. K. Warfield, “Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE,” *IEEE Trans. Med. Imaging* **31**(8), 1593–1606 (2012).

²⁷M. R. Sabuncu *et al.*, “A generative model for image segmentation based on label fusion,” *IEEE Trans. Med. Imaging* **29**(10), 1714–1729 (2010).

²⁸R. Wolz *et al.*, “Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012*, Nice, France (Springer, Berlin, Heidelberg, 2012), pp. 10–17.

²⁹A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc., Ser. B* **39**, 1–38 (1977).

- ³⁰T. Rohlfing, D. B. Russakoff, and C. R. Maurer, "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *IEEE Trans. Med. Imaging* **23**(8), 983–994 (2004).
- ³¹A. J. Asman *et al.*, "Robust GM/WM segmentation of the spinal cord with iterative non-local statistical fusion," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2013* (Springer, Nagoya, Japan, 2013).
- ³²B. M. Dawant *et al.*, "Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations. I. Methodology and validation on normal subjects," *IEEE Trans. Med. Imaging* **18**(10), 909–916 (1999).
- ³³J. M. Lee *et al.*, "Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error," *Comput. Biol. Med.* **33**(6), 495–507 (2003).
- ³⁴A. Evans *et al.*, "MRI-PET correlation in three dimensions using a volume-of-interest (VOI) atlas," *J. Cereb. Blood Flow Metab.* **11**, A69–A78 (1991).
- ³⁵M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.* **6**(2), 181–214 (1994).
- ³⁶D. Windridge and J. Kittler, "A morphologically optimal strategy for classifier combination: Multiple expert fusion as a tomographic process," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(3), 343–353 (2003).