Simultaneous Truth and Performance Level Estimation with Incomplete, Over-complete, and Ancillary Data

Bennett A. Landman[*a,c], John A. Bogovic[b], Jerry L. Prince[a,b]
[a]Biomedical Engineering, [b]Electrical and Computer Engineering,
Johns Hopkins University, 3400 N. Charles St., Baltimore, MD, USA 21218
[c]Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235

ABSTRACT

Image labeling and parcellation are critical tasks for the assessment of volumetric and morphometric features in medical imaging data. The process of image labeling is inherently error prone as images are corrupted by noise and artifact. Even expert interpretations are subject to subjectivity and the precision of the individual raters. Hence, all labels must be considered imperfect with some degree of inherent variability. One may seek multiple independent assessments to both reduce this variability as well as quantify the degree of uncertainty. Existing techniques exploit maximum *a posteriori* statistics to combine data from multiple raters. A current limitation with these approaches is that they require each rater to generate a complete dataset, which is often impossible given both human foibles and the typical turnover rate of raters in a research or clinical environment. Herein, we propose a robust set of extensions that allow for missing data, account for repeated label sets, and utilize training/catch trial data. With these extensions, numerous raters can label small, overlapping portions of a large dataset, and rater heterogeneity can be robustly controlled while simultaneously estimating a single, reliable label set and characterizing uncertainty. The proposed approach enables parallel processing of labeling tasks and reduces the otherwise detrimental impact of rater unavailability.

Keywords: Parcellation, labeling, delineation, statistics, data fusion, analysis, STAPLE

1. INTRODUCTION

Numerous clinically relevant conditions (e.g., degeneration, inflammation, vascular pathology, traumatic injury, cancer, etc.) correlate with volumetric/morphometric features as observed on magnetic resonance imaging (MRI). Quantification and characterization of these correlations requires the labeling or delineation of structures of interest. The established gold standard for identifying class memberships is manual voxel-by-voxel labeling by a neuroanatomist, which can be exceptionally time and resource intensive. Furthermore, different human experts often have differing interpretations of ambiguous voxels (on the order of 5-10% of a typical brain structure). Therefore, pursuit of manual approaches is typically limited to either (1) validating automated or semi-automated methods or (2) the study of structures for which no automated method exists.

An often understood objective in manual labeling is for each rater produce the most accurate and reproducible labels possible. Yet, this is not the only possible technique for achieving reliable results. Kearns and Valiant first posed the question whether a collection of "weak learners" (raters that are just better than chance) could be boosted ("combined") to form a "strong learner" (a rater with arbitrarily high accuracy) [1]. The first affirmative response to this challenge was proven a year later [2]. With the presentation of AdaBoost, boosting became widely practical [3].

Statistical methods have been previously proposed to simultaneously estimate rater reliability and true labels from complete datasets created by several different raters or automated methods [4-7]. While there are typically many fewer raters available in brain imaging research, and raters generally are considered superior to "weak learners." Warfield et al. presented a probabilistic algorithm to estimate the "ground truth" segmentation from a group of expert segmentations and simultaneously assess of the quality of each expert [4]. Rohlfing et al. also employed this approach to multiple labels [6]. These maximum likelihood/maximum *a posteriori* methods (e.g., Simultaneous Truth and Performance Level Estimation, STAPLE [5]) increase the accuracy of a single labeling by combining information from multiple, potentially less accurate raters (as long as the raters are independent and collectively unbiased). However, the existing methods

* bennett.landman@vanderbilt.edu; http://masi.vuse.vanderbilt.edu, http://iacl.ece.jhu.edu; Image Analysis and Communications Laboratory, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA 21218

Medical Imaging 2010: Image Processing, edited by Benoit M. Dawant, David R. Haynor,
Proc. of SPIE Vol. 7623, 76231N · © 2010 SPIE · CCC code: 1605-7422/10/$18 · doi: 10.1117/12.844182

require that all raters delineate all voxels, which limits applicability in real research studies where different sets of raters may delineate arbitrary subsets of a population of scans due to the rater availability or the scale of the study.

Herein, we present and demonstrate Simultaneous Truth and Performance Level Estimation with Robust extensions (STAPLER) to enable use of data with:

1. **Missing labels**: partial labels sets in which raters do not delineate all voxels;
2. **Repeated labels**: labels sets in which raters may generate repeated labels for some (or all) voxels; and
3. **Training trials**: label sets in which some raters may have known reliabilities (or some voxels have known true labels). These may also be derived from catch trials. We consider this information ancillary as it does not specifically relate to the labels on structures of interest, but rather to the variability of individual raters.

STAPLER simultaneously incorporates all labels from all raters to estimate a maximum *a posteriori* estimate of both rater reliability and true labels. The impacts of missing and training data are studied with simulations based on two models of rater behavior. First, the performance is studied using traditional "random raters," which are parameterized by confusion matrices (i.e., probabilities of indicating each label given a true label). Second, we develop a new, more realistic set of simulations in which raters make more mistakes along the boundaries between regions. The performance of STAPLER is characterized with these simulated rater models in simulations of cerebellar parcellation.

2. METHODS

STAPLE exploits expectation maximization to calculate rater reliabilities, Θ_{jsT}^{k}, i.e., the probability that a rater, j, reports that a voxel, i, has a particular label, s, given a true label, T. Rater reliabilities and observed data, D_{ijr}, with r repetitions can be used to calculate the conditional probability that a voxel belongs to a class, W_{si}^{k}, at iteration k. In [5], the conditional expectation of the complete data log likelihood is reported as (for all raters reporting at all voxels, Eq. 20):

$$W_{si}^{k} = p\left(T = s \middle| \boldsymbol{D_i}, \boldsymbol{\Theta^k}\right) = \frac{p(T_i=s)\prod_j \Theta_{jsT}^{k}}{\sum_{s'} p(T_i=s')\prod_j \Theta_{jsT}^{k}} \tag{1}$$

When this formulation is extended to include contributions from observed data, product terms adjust to exclude unobserved data points:

$$W_{si}^{k} = p\left(T = s \middle| \boldsymbol{D_i}, \boldsymbol{\Theta^k}\right) = \frac{p(T_i=s)\prod_{j:D_{ijr}\neq\emptyset} \Theta_{jsT}^{k}}{\sum_{s'} p(T_i=s')\prod_{j:D_{ijr}\neq\emptyset} \Theta_{jsT}^{k}} \tag{2}$$

Second, in [5], the update equation for parameter estimates was derived as (for all rater reporting at all voxels and with no "known" data, Eq. 24):

$$\Theta_{jsT}^{k+1} = \frac{\sum_{i:D_{ij}=s} W_{Ti}^{k}}{\sum_i W_{Ti}^{k}} \tag{3}$$

where I is the indicator function. To extend in this framework to the STAPLER case, we perform three modifications. First, parameters for raters with known reliabilities are not updated. Second, if a true label set is given, then an additional rater is introduced and modeled as if that rater reported the true labels. This rater is modeled as a rater with known reliability equal to one. Third, the update equation for the remaining raters is generalized to include contributions from all available data. In summary,

$$\begin{cases} \Theta_{jsT} \text{ fixed} \rightarrow \text{ no update} \\ 0 = \sum_{i:D_{ijr}=s} W_{Ti}^{k} \rightarrow \Theta_{jsT}^{k+1} = I\{s=T\} \\ \text{otherwise} \rightarrow \Theta_{jsT}^{k+1} = \frac{\sum_{i:D_{ijr}=s} W_{Ti}^{k}}{\sum_{i:D_{ijr}\neq\emptyset} W_{Ti}^{k}} \end{cases} \tag{4}$$

where I is the indicator function.

There are several possible routes ones could take to model the unconditional label probabilities (i.e., the label priors). If the relative sizes of the structures of interest are known, a fixed probability distribution could be used. Alternatively, one could employ a random field model to identify probable points of confusion (as in [5]). The simpler models have the potential for introducing unwanted bias while field based models may suffer from slow convergence. Herein, we use an adaptive mean label frequency to update the unconditional label probabilities:

$$p(T_i = s)^{k+1} = \frac{\sum_i \mathrm{W}_{jsT}^k}{\sum_{iT} \mathrm{W}_{jsT}^k} \tag{5}$$

STAPLER was implemented in Matlab (Mathworks, Natick, MA). A custom toolbox provided efficient access to large sparse matrices. All studies were run on a 64 bit 2.5 GHz notebook with 4 GB of RAM. As in [5], simultaneous parameter and truth level estimation was performed with iterative expectation maximization.

Experiments with random raters were performed with a known, true ground truth model. The accuracy of each label set (either from an individual or reconstructed with label fusion, A) was assessed relative to the truth model (B) with the Jaccard similarity index [8, 9] for each labeled region:

$$\boldsymbol{J_T(A,B)} = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A} \cup \boldsymbol{B}|} = \frac{\sum_{\boldsymbol{A_i=T \text{ and } B_i=T}} \mathbf{1}}{\sum_{\boldsymbol{A_i=T \text{ or } B_i=T}} \mathbf{1}} \tag{6}$$

The Jaccard index ranges from 0 (indicating no overlap between label sets) to 1 (indicating no disagreement between label sets).

3. DATA

Imaging data were acquired from two healthy volunteers who provided informed written consent prior to study. A high resolution MPRAGE (magnetization prepared rapid acquired gradient echo) sequence was acquired axially with full head coverage (149x81x39 voxels, 0.82x0.82x1.5 mm resolution). An experienced human rater labeled the cerebellum from each dataset with 12 divisions of the cerebellar hemispheres (Figure 3A/1B) [10, 11]. Simulated label sets were derived from simulated raters using a Monte Carlo framework.

Two distinct models of raters (described below) were evaluated as follows:

1. Random raters were simulated: *Rater characteristics were generated through pseudo-randomization of the given performance model.*
2. Simulated label sets from the raters were generated according to the profiles: *These datasets corresponded to synthetic labelings of the two MRI datasets given the performance characteristic of each rater.*
3. Traditional STAPLE was evaluated by combining labels from 3 random raters. *Each of the three synthetic raters was modeled as having labeled one complete dataset.*
4. STAPLER was evaluated by labels from 3*M raters where 3 raters were randomly chosen to delineate each slice. *Each rater delineated approximately 1/Mth (i.e., each rater labels between 50% and 4% of slices with the total amount of data held constant).*
5. The advantages of incorporating training data were studied by repeating the STAPLER analysis with all raters also fully labeling a second, independent test data set with known true labels.

Note that in the case of M=1 and the absence of training data, STAPLER is equivalent to STAPLE. The procedure was repeated either 10 or 25 times (as indicated below) and the mean and standard deviation of overlap indices were reported for each analysis method.

3.1 Traditional Random Raters (errors distributed evenly within the volume)

In the first model (Figure 1), each rater was assigned a confusion matrix such that the *i,j*th element indicates the probability that the rater would assign the *j*th label when the *i*th label is correct. Label errors are equally likely to occur throughout the image domain and exhibit no spatial dependence. The background region is considered a labeled region. This is the same model of rater performance as employed by the statistical framework.

To generate each pseudo-random rater, a matrix with each entry corresponding to a uniform random number between 0 and 1 was created. The confusion matrix was generated by adding a scaled identity matrix to the randomly generated matrix and normalizing column sums to one such that the mean probability of true labels was 0.93 (e.g., the mean diagonal element was 0.93). Ten Monte Carlo iterations were used for each simulation.

3.2 New, Boundary Random Raters (errors distributed along label boundaries)

In the second model (Figure 2), rater errors occurred at the boundaries of labels rather than uniformly throughout the image domain. Three parameters describe rater performance: r, $\boldsymbol{l}$, and $\boldsymbol{b}$. The scalar r is the rater's global true positive fraction. The boundary probability vector $\boldsymbol{l}$ encodes the probability, given an error occurred, that it was at the ith boundary. Finally the vector $\boldsymbol{b}$ describes the error bias at every boundary which denotes the probability of shifting a boundary toward either bounding label. For an unbiased rater, $\boldsymbol{b}_i = 0.5, \forall i$. Twenty-five Monte Carlo iterations were used for each simulation.

To generate a pseudo-random rater, the boundary probability vector was initialized to a vector with uniform random coefficients and normalized to sum to 1. To generate a simulated random dataset with a given boundary rater, the voxel-wise mask of truth labels was first converted into a set of boundary surfaces. Then, the following procedure was repeated for $(1 - r)$ |B| iterations (where N is the set of all image voxels).

1. A boundary surface (a pair of two labels) was chosen according to the $\boldsymbol{l}$ distribution. If the boundary did not exist in the current dataset, a new boundary surface was chosen until it did exist.
2. A boundary point within the chosen surface was selected uniformly at random for all boundary points between the two label sets.
3. A random direction was chosen Bernoulli($\boldsymbol{b}_i$) to determine if the boundary surface would move toward label pair with the lower index or the label pair with the high index.
4. The set of boundary voxels was updated to reflect the change in boundary position. With the change in labels, the set of boundary label boundary pairs was also updated since the changes in voxel classification can lead to changes in the topology of the surface collections.

In this study, the rater performance was set to 0.8 and the bias term was set to 0.5. The boundary random rater framework was implemented in the Java Image Science Toolkit (JIST, http://www.nitrc.org/projects/jist/ [12, 13]).

4. RESULTS

4.1 Traditional Random Raters

For a single rater, the Jaccard index was 0.67±0.02 (mean ± standard error over simulated datasets, one label set is shown in Figure 3C). The traditional STAPLE approach with three raters visually improved the consistency of the results (one label set is shown in Figure 3D); the average Jaccard index with STAPLE also increased to 0.98±0.012 (first column of Figure 3E). For all STAPLER simulations, use of multiple raters improved the label reliability over that which was achievable with single rater (Figure 3E).

STAPLER consistently resulted in Jaccard indexes above 0.9, even when each individual rater labeled 10 percent of the dataset. While the Jaccard index was equivalent to that of the STAPLE approach when raters labeled as little as one third of the dataset, the achievable consistency with less overlap resulted in appreciably degraded performance. The decrease in reliability arises because not all raters have observed all labels with equal frequency. For smaller regions, some raters may have observed very few (or no data points). During estimation, the rater reliabilities for these "under seen" labels are very noisy and led to unstable estimates, which result in estimation of substantial off-diagonal components of the confusion matrix. Note that all simulations were designed such that each voxel was labeled exactly three times; only the identity of the simulated rater who contributed these labels varied.

Use of training trials greatly improved the accuracy of label estimation when many raters each label a small portion of the data set (Figure 3E). No appreciable differences were seen when the number of raters were varied. The use of training data effectively places a data-adaptive prior on the confusion matrix. Since each rater provides a complete dataset, each label category is observed by each rater for a substantial quantity of voxels. Hence, the training data provide evidence against artifactual, large off-diagonal confusion matrix coefficients and improves estimation stability. Furthermore, without missing categories, there are no undetermined confusion matrix entries.

4.2 New, Boundary Random Raters

For a single rater, the Jaccard index was 0.83±0.01 (one label set shown in Figure 4B). Using three raters in a traditional STAPLE approach increased the average Jaccard index to 0.91±0.01 (one label set shown in Figure 4E). The STAPLER approach lead to consistently high Jaccard indexes with as low as 25 percent of the total dataset labeled by each rater. However, with individual raters generating very limited data sets (<10%), STAPLER was prone to "label inversion" and resulted *increased error* over a single rater. In this case, off-diagonal elements of the estimated confusion matrix become large and lead to label switching (Figure 4C,E-G). This behavior was not routinely observed in the first experiment, but was one factor that led toward increased variability of Jaccard index (see outlier data points in Figure 3E). As in the first experiment, all simulations were designed such that each voxel was labeled exactly three times; only the identity of the simulated rater who contributed these labels varied.

Use of data from training trials alleviates this problem by ensuring that sufficient data on each label from each rater is available (Figure 4D,H-J). The Jaccard index showed no appreciable differences when raters labeled between 4 percent and 100 percent of the dataset. However, the use of training data stabilized the reliability matrix estimation process. The artifactual, large off-diagonal confusion matrix coefficients were not observed when training data were used.

5. CONCLUSIONS

STAPLER extends the applicability of the STAPLE technique to common research situations with missing, partial, and repeated data and facilitates use of training data to improve accuracy. These ancillary data are commonly available and may either have exact known labels or be labeled by a rater with known reliability. A typical scenario would involve a period of rater training followed by their carrying out a complete labeling on the training set. Only then would they carry out independent labeling of test data. STAPLE was successful both when simulated error matched modeled errors (i.e., the traditional model) and with more realistic, boundary errors, which is promising for future application to work involving efforts of large numbers of human raters.

The traditional approach to manually label images has been pixel-by-pixel annotation by a rater or team of raters. Yet, expert raters are a very limited resource given their extensive anatomical and imaging understanding; an experienced neurologist requires nearly a decade of training. Individuals with a biological background can achieve reasonable reliability for specific structures in much less time (typically within approximately 3 months training for our cerebellar protocols). However, well qualified individuals tend to quickly move on in search of more varied work. Historically, use of trained (but non-career track) labelers has been plagued by difficulties in achieving long-term consistency and reliability.

With the newly presented technique, numerous raters can label small, overlapping portions of a large dataset, which can be recombined into a single, reliable label estimate, and the time commitment from any individual rater can be minimized. This enables parallel processing of manual labeling and reduces detrimental impacts should a rater become unavailable during a study. Hence, less well trained raters or raters who may participate on a part-time basis could contribute. As with STAPLE, both the labels and degrees of confidence on those labels are simultaneously estimated, so that subsequent processing could make informed decisions regarding data quality. Such an approach could enable collaborative image labeling system and be a viable alternative to expert raters in neuroscience research.

Evaluation of STAPLER with partially labeled datasets from human raters is an active area of research and will be reported in subsequent publications. The improvement in Jaccard index in the boundary rater model was less than that in the traditional random rater model (from 0.83 to 0.91 versus 0.67 to 0.98). In the traditional rater example, both the estimation and underlying error models were the same. In the boundary rater model, the model used during estimation was only a loose approximation of the underlying mechanism. This result provides an indication that simple rater confusion models may still be effective in practice (with human rater) when difficult to characterize interdependencies might exist between rater confusion characteristics, the data, and temporal characteristics. Additionally, it hints that more comprehensive models might enable more efficient label fusion from human raters. As with the original STAPLE algorithms, STAPLER can readily be improved by introducing spatially adaptive unconditional label probabilities, such as with a Markov Random Field (MRF). STAPER extensions are independent of the manifold of the underlying data. These methods are equally applicable to fusion of volumetric labels [14-16], labeled surfaces[17, 18], or other point-wise structures.

ACKNOWLEDGEMENTS

This project was supported by NIH/NINDS 1R01NS056307 and NIH/NINDS 1R21NS064534.

This work described herein has not been submitted elsewhere for publication or presentation. It is an expansion of work previously published in abstract form at the 2009 International Society of Magnetic Resonance in Medicine conference [19].

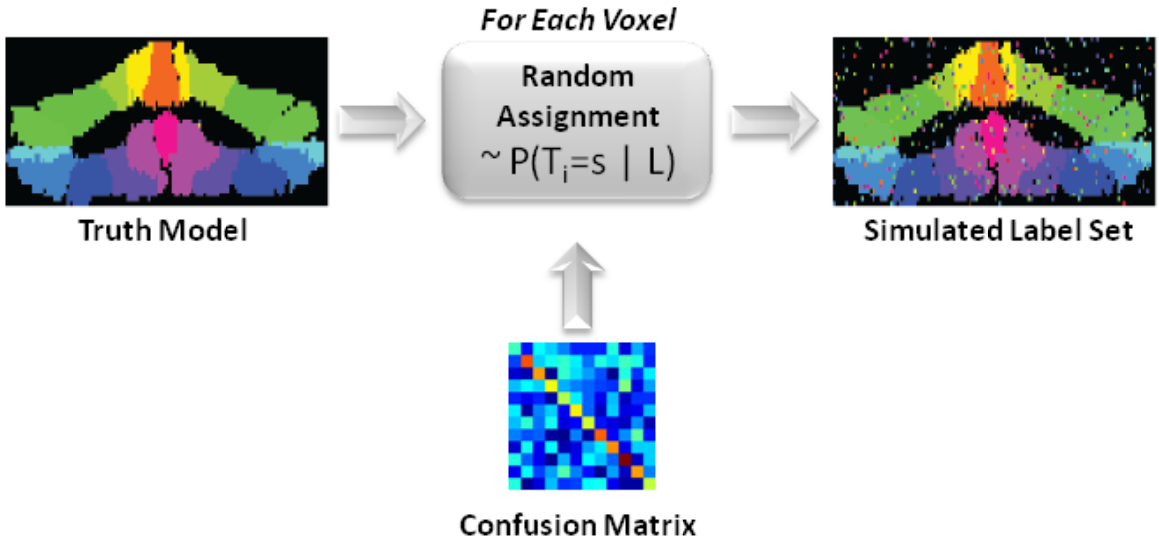

Figure 1. Traditional random rater model. The distribution of label probabilities depends on the underlying true label, but does not depend on the local neighborhood or spatial position.

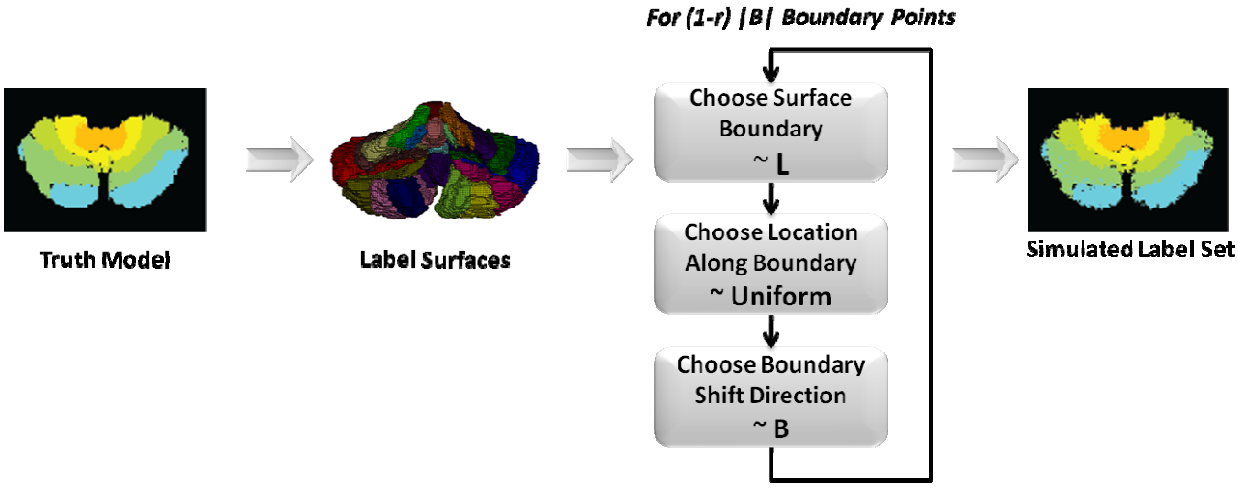

Figure 2. Boundary random rater model. Errors are uniformly distributed on the boundaries between regions. Sampling of boundary errors is done iterative with replacement and model updating so that it is possible for cumulative errors to shift the boundary by multiple voxels in any location. Boundary surfaces are stored at voxel resolution on a Cartesian grid.

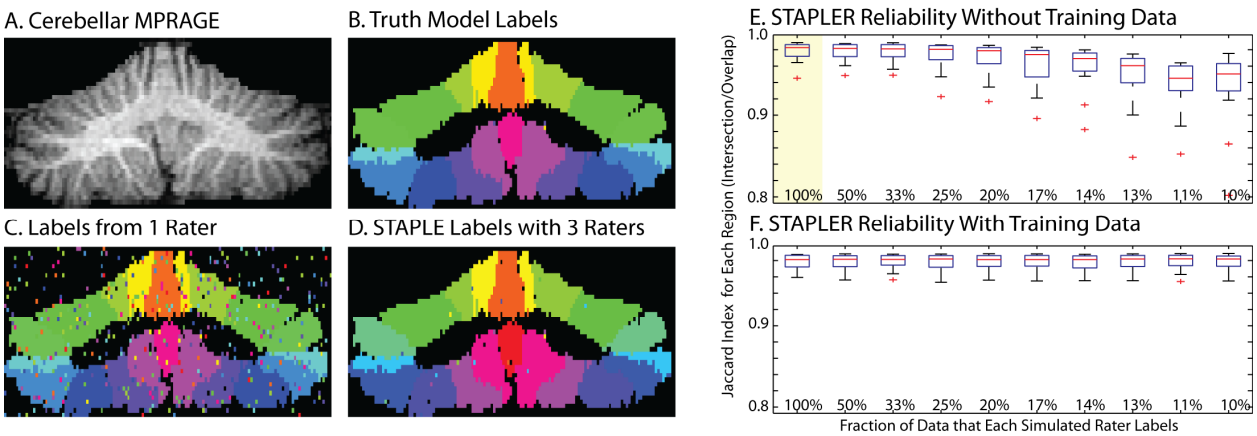

Figure 3. Simulations with traditional random raters. Coronal sections of the three-dimensional volume show the high resolution MRI image (A), manually drawn truth model (B), an example delineation from one random traditional rater (C), and the results of a STAPLE recombination of three label sets (D). STAPLER enables fusion of label sets when raters provide only partial datasets, but performance suffers with decreasing overlap (E). With training data (F), STAPLER improved the performance even with each rater labeling only a small portion of the dataset. Box plots in E and F show mean, quartiles, range up to 1.5σ, and outliers. The highlighted plot in E indicates the simulation for which STAPLER was equivalent to STAPLE--i.e., all raters provide a complete set of labels.

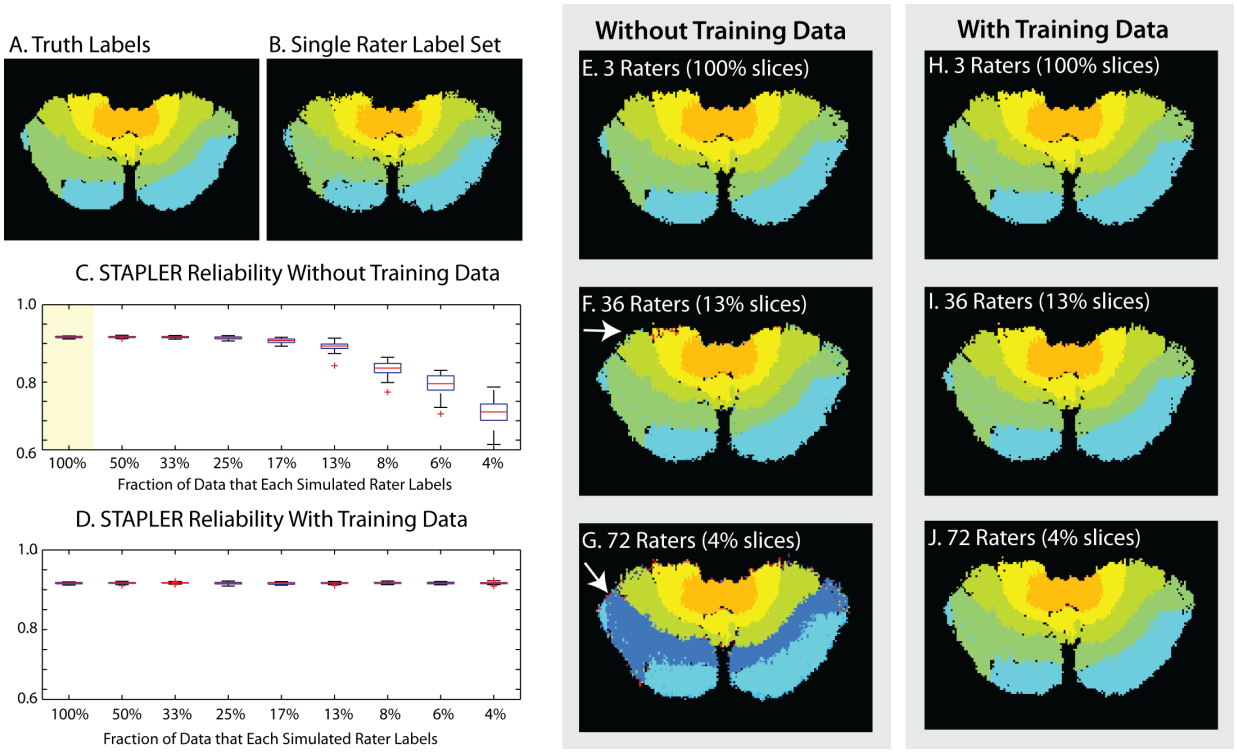

Figure 4. Simulations with boundary random raters. Axial sections of the three-dimensional volume show the manually drawn truth model (A) and sample labeling from a single simulated rater (B) alongside STAPLER fused results from 3, 36, and 72 raters producing a total of 3 complete labeled datasets without training data (E-G) and with training data (H-J). Note that boundary errors are generated in three-dimensions, so errors may appear distant from the boundaries in cross-sections. Boundary errors (e.g., arrow in F) increased with decreasing rater overlap. Label inversions (e.g., arrow in G) resulted in very high error with minimal overlap. As with the traditional model (Figure 1), STAPLER enables fusion of label sets when raters provide only partial datasets, but performance suffers with decreasing overlap (C). With the addition of training data (D), STAPLER results in sustained performance improvement even with each rater labeling only a small portion of the dataset.

REFERENCES

[1] M. Kearns, and L. G. Valiant, “Learning boolean formulae or finite automata is as hard as factoring,” Harvard University Technical Report, TR-14-88, (1998).

[2] R. E. Shapire, “The strength of weak learnability,” Machine Learning, 5(2), 197-227 (1990).

[3] Y. Freund, and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” J. Computer and System Sciences, 55(119–139), (1997).

[4] S. K. Warfield, K. H. Zou, M. R. Kaus, and W. M. Wells, [Simultaneous validation of image segmentation and assessment of expert quality], Washington, DC(2002).

[5] 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 Trans Med Imaging, 23(7), 903-21 (2004).

[6] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, “Expectation maximization strategies for multi-atlas multi-label segmentation,” Inf Process Med Imaging, 18, 210-21 (2003).

[7] J. Udupa, V. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. Currie, B. Hirsch, and J. Woodburn, “A framework for evaluating image segmentation algorithms,” Comp Med Imag Graphics 30(2), 75-87 (2006).

[8] J. C. Gee, M. Reivich, and R. Bajcsy, “Elastically deforming 3D atlas to match anatomical brain images,” J. Comput. Assist. Tomogr., 17, 225–236 (1993).

[9] P. Jaccard, “The distribution of flora in the alpine zone,” New Phytol. , 11, 37–50 (1912).

[10] N. Makris, S. M. Hodge, C. Haselgrove, D. N. Kennedy, A. Dale, B. Fischl, B. R. Rosen, G. Harris, V. S. Caviness, and J. D. Schmahmann, “Human cerebellum: surface-assisted cortical parcellation and volumetry with magnetic resonance imaging,” J Cogn Neurosci, 15(4), 584-599 (2003).

[11] N. Makris, J. Schlerf, S. Hodge, C. Haselgrove, M. Albaugh, L. Seidman, S. Rauch, G. Harris, J. Biederman, V. Caviness, D. Kennedy, and J. Schmahmann., “MRI-based surface-assisted parcellation of human cerebellar cortex: an anatomically specified method with estimate of reliability,” Neuroimage, 25(4), 1146-1160 (2005).

[12] B. C. Lucas, J. A. Bogovic, A. Carass, P.-L. Bazin, J. L. Prince, D. Pham, and B. A. Landman, “The Java Image Science Toolkit (JIST) for Rapid Prototyping and Publishing of Neuroimaging Software,” Neuroinformatics, (In press 2010).

[13] B. A. Landman, B. C. Lucas, J. A. Bogovic, A. Carass, and J. L. Prince, "A Rapid Prototyping Environment for NeuroImaging in Java."

[14] A. Dimitrova, D. Zeljko, F. Schwarze, M. Maschke, M. Gerwig, M. Frings, A. Beck, V. Aurich, M. Forsting, and D. Timmann, “Probabilistic 3D MRI atlas of the human cerebellar dentate/interposed nuclei,” Neuroimage, 30(1), 12-25 (2006).

[15] B. A. Landman, A. X. Du, W. D. Mayes, J. L. Prince, and S. H. Ying, "Diffusion Tensor Imaging Enables Robust Mapping of the Deep Cerebellar Nuclei."

[16] J. Bogovic, B. Landman, J. Prince, and S. Ying, "Probabilistic Atlas of Cerebellar Degeneration Reflects Volume and Shape Changes."

[17] J. A. Bogovic, A. Carass, J. Wan, B. A. Landman, and J. L. Prince, [Automatically identifying white matter tracts using cortical labels], Paris, France(2008).

[18] J. Bogovic, B. A. Landman, P.-L. Bazin, and J. L. Prince, "Statistical Fusion of Surface Labels provided by Multiple Raters, Over-complete, and Ancillary Data."

[19] B. A. Landman, J. Bogovic, and J. L. Price, "Efficient Anatomical Labeling by Statistical Recombination of Partially Label Datasets."