# Efficient Anatomical Labeling by Statistical Recombination of Partially Label Datasets

**B. A. Landman[1], J. A. Bogovic[2], and J. L. Prince[2,3]**

[1]Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States, [2]Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, United States, [3]Biomedical Engineering, Johns Hopkins University, Balitmore, MD, United States

**Introduction:** Numerous clinically relevant conditions (e.g., degeneration, inflammation, vascular pathology, tramatic injury, cancer, etc.) correlate with volumetric/morphometric changes as observed on MRI. Study of these changes necessitates the ability to label (delineate) voxel-wise classifications for structures of interest. The established gold standard for identifying class memberships is manual voxel-by-voxel labeling by a neuro-anatomist expert, 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. Previously, statistical methods have been presented to simultaneously estimate rater reliability and true labels from complete datasets by several different raters [1-3]. These maximum likelihood/maximum *a posteriori* methods (i.e., STAPLE) increase the accuracy of a single labeling by combining information from multiple, potentially less accurate raters (so long as the set of raters is independent and unbiased). However, the existing methods 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 duration of the study. Here, we present 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 (independent) labels for some (or all) voxels; and (3) **Catch trials**: label sets in which some raters may have known reliabilities (or some voxels have known true labels). STAPLER simultaneously incorporates all labels from all raters to estimate a maximum *a posteriori* estimate of both rater reliability and labels.

**Methods:** The STAPLE methods [1-3] use expectation maximization to calculate the rater reliabilities ($\Theta_{jsT}^{k}$), 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 repetition $r$ can be used to calculate the conditional probability that a voxel belongs to a class ($W_{si}^{k}$) at iteration $k$. First, we extend Eq. 20 in [1] to include all observed data:

$$W_{si}^{k} = p(T = s \,|\, \boldsymbol{D}_i, \boldsymbol{\Theta^k}) = \frac{p(T_i=s)\prod_{j:D_{ijr}\neq\emptyset}\Theta_{jsT}^{k}}{\sum_{s'}p(T_i=s')\prod_{j:D_{ijr}\neq\emptyset}\Theta_{js'T}^{k}}$$

Second, we extend Eq. 24 in [1] to prevent update of rater reliabilities for raters with known reliabilities or no observed data:

If $\Theta_{jsT}$ is fixed, then no update. If $0=\sum_{i:D_{ijr}=s} W_{Ti}^{k}$, then $\Theta_{jsT}^{k+1} = I\{s = T\}$, where $I$ is the indicator function. Otherwise: $\Theta_{jsT}^{k+1} = \frac{\sum_{i:D_{ijr}=s} W_{Ti}^{k}}{\sum_{i:D_{ijr}\neq\emptyset} W_{Ti}^{k}}$.

If a subset of truth labels is given, then an additional rater is introduced for these voxels with known perfect reliability. STAPLER was implemented in Matlab (Mathworks, Natick, MA) using a custom sparse matrix toolbox and evaluated on a 2.5 GHz 64 bit Linux workstation. Unconditional label probabilities ($p(T_i = s)$) can be defined by the user. Here, we use a global adaptive mean. Initialization was performed with equal probabilities for all labels.

To demonstrate the utility of this approach, we generate random, simulated raters based on a high resolution labeling of 12 divisions of the cerebellar hemispheres (**Fig. 1&2**) (149x81x39 voxels, 0.821x0.82x1.5 mm resolution). For clarity, coronal sections are shown; however, data were acquired and analyzed on an axial basis. Each rater was randomly assigned a reliability matrix such that the average true positive rate was 0.93. The following scenarios were each studied using 10 Monte Carlo iterations: (1) Labels from individual raters were evaluated by generating labels according to the rater reliability profiles; (2) Traditional STAPLE was evaluated by combining labels from 3 random raters; (3) STAPLER was evaluated by labels from 3*M raters where 3 raters were randomly chosen to delineate each slice, and each rater delineated approximately $1/M^{th}$ of the dataset for M=2-10 (i.e., each rater labels between 50% and 10% of slices); (4) The advantages of incorporating catch trials was studied by repeating experiment 3 with M =1-10 and having all raters fully label a second, independent test data set with known labels.

**Results:** The expected Jaccard index (i.e., intersection divided by union) for a single rater was 0.67+-0.02 (range 0.22-0.86) (one label set shown in **Fig. 3**). Using three raters in a traditional STAPLE approach increased the average Jaccard index to 0.98+-0.012 (range 0.978-0.981) (one label set shown in **Fig. 4**, estimation time=43 s). Although STAPLER provide high reliability statistical recombination of sparse label sets even when each rater delineates 10% of the total data (**Fig. 5A**), performance degrades with decreasing overlap. The decrease in reliability arises because not all raters have observed all labels with equal frequency, so the rater reliabilities for the unseen labels are under-determined and lead to unstable estimates. Use of catch trials (e.g., from training data) greatly improve the accuracy of label estimation when many raters each label a small portion of the data set (**Fig. 5B**). Box plots show mean, quartiles, range up to 1.5σ, and outliers.

**Discussion:** STAPLER extends the applicability of the STAPLE technique to common research situations with missing, partial, and repeated data challenges. Furthermore, STAPLER facilitates use of training data to improve labeling accuracy. Evaluation with human raters is underway. These supplementary data are commonly available in practices and may either have exact known labels or be labeled by a rater with known reliability. With the newly presented technique, numerous raters can label small, overlapping portions of a large dataset, which can then 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. 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). Simple, yet effective, volumetric MRF models may be introduced with nominal memory and computational overhead.

**References:** [1] Warfield, SK, Zou, KH, and Wells, WM. IEEE TMI. 2004. 23(7): 903 [2] Rohlfing, T, Russakoff, DB, Maurer, CR. IEEE TMI. 2004. 23(8):983 [3] Udupa, JK, LeBlanc, et al. Comp Med Imag Graphics. 2006. 30(2):75

Fig. 1. Cerebellar MPRAGE

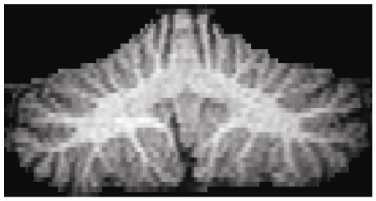

Fig. 2. Truth Model Labels

Fig. 3. Labels from 1 Rater

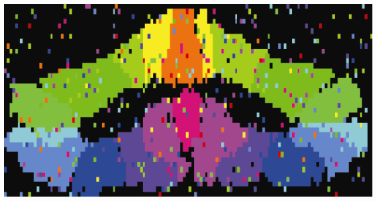

Fig. 4. STAPLE Labels with 3 Raters

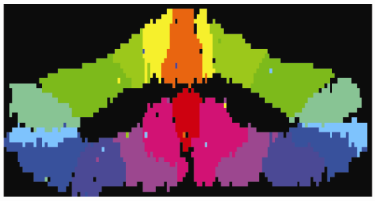

Fig. 5. STAPLER Reliability with 3 Raters for Each Slice