Published in final edited form as:
*Proc Soc Photo Opt Instrum Eng*. 2011 ; 7962: 79623G. doi:10.1117/12.877471.

# Foibles, Follies, and Fusion: Assessment of Statistical Label Fusion Techniques for Web-Based Collaborations using Minimal Training

**Andrew J. Asman**[a,*], **Andrew G. Scoggins**[a], **Jerry L. Prince**[b], and **Bennett A. Landman**[a,c]

[a] Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235

[b] Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218

[c] Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218

## Abstract

Labeling or parcellation of structures of interest on magnetic resonance imaging (MRI) is essential in quantifying and characterizing correlation with numerous clinically relevant conditions. The use of statistical methods using automated methods or complete data sets from several different raters have been proposed to simultaneously estimate both rater reliability and true labels. An extension to these statistical based methodologies was proposed that allowed for missing labels, repeated labels and training trials. Herein, we present and demonstrate the viability of these statistical based methodologies using real world data contributed by minimally trained human raters. The consistency of the statistical estimates, the accuracy compared to the individual observations and the variability of both the estimates and the individual observations with respect to the number of labels are discussed. It is demonstrated that the Gaussian based statistical approach using the previously presented extensions successfully performs label fusion in a variety of contexts using data from online (Internet-based) collaborations among minimally trained raters. This first successful demonstration of a statistically based approach using "wild-type" data opens numerous possibilities for very large scale efforts in collaboration. Extension and generalization of these technologies for new application spaces will certainly present fascinating areas for continuing research.

### Keywords

Parcellation; labeling; delineation; label fusion; STAPLE; STAPLER; minimal training

## 1. INTRODUCTION

Labeling or delineation of structures of interest on magnetic resonance imaging (MRI) is essential in quantifying and characterizing correlation with numerous clinically relevant conditions. These conditions include but are not limited to degeneration, inflammation, traumatic injury, cancer and vascular pathology. The established standard for delineation and segmentation of MRI is manual voxel-by-voxel labeling by a neuroanatomist. This process can be extremely time consuming, resource intensive, and fraught with variability. Different human experts may have disagreements at ambiguous pixels (composing 5–10% of a typical brain structure) and their decisions would be based upon individual

[*]andrew.j.asman@vanderbilt.edu; http://masi.vuse.vanderbilt.edu; Medical-image Analysis and Statistical Interpretation Laboratory, Department of Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235.

interpretation. Thus, the scope of manual approaches is typically limited to (1) validating automated or semi-automated methods or (2) the study of structures for which no automated method exists [10].

The most commonly understood objective in manual labeling is that each rater should produce the most accurate and reproducible labels possible. However, this is not the only technique for achieving the highest possible results. The technique proposed by Kearns and Valiant suggested that 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 suggestion was proven a year later [2]. With the presentation of AdaBoost, boosting became widely practical [3].

The use of statistical methods using automated methods or complete data sets from several different raters have been proposed to simultaneously estimate both rater reliability and true labels [4–7]. Warfield et al. presented an algorithm that used probability to estimate the “ground truth” segmentation from a group of expert segmentations and simultaneously assess the quality of each expert [4]. The approach was extended by Rohlfing et al. [6]. These methods are based upon a maximum likelihood/maximum *a posteriori* approach (e.g. Simultaneous Truth and Performance Level Estimation, STAPLE [5]) and they increase the accuracy of a single labeling by probabilistically combining multiple, potentially less accurate segmentations (assuming that the raters are independent and collectively unbiased). Nevertheless these approaches require each rater to delineate all voxels, which limits the viability of these approaches for real world research studies.

Extensions to these statistical based methodologies were proposed that allowed for missing labels, repeated labels and training trials [10–12]. Additionally, this extension provided a method of using a biasing prior that decreases the susceptibility to outlier delineations. This method (Simultaneous Truth and Performance Level Estimation with Robust extensions (STAPLER) [10]) made significant advancements to previous statistical algorithms and increased the viability of using these algorithms on real world data sets.

Herein, we present and demonstrate the use of these statistical based methodologies using real world data contributed by minimally trained human raters. The algorithm is demonstrated to work in situations where the foibles and follies of human raters are present and included in the data set used to construct the final “ground truth” estimation. Three main data sets are presented in this paper (1) pig cardiac label set, (2) simulated label set (spheres of various radii placed randomly in a cylinder) and (3) axial, sagittal and coronal cross sections of a cerebellum. These data sets range from relatively simple labeling (pig cardiac data set, simulated data set) to significantly more challenging label sets (axial, sagittal and coronal cross sections of the cerebellum). The consistency of STAPLE estimates, the accuracy compared to the individual observations and the variability of both the estimates and the individual observations with respect to the number of labels are discussed. Additionally, the accuracy disparity between the training and testing data sets and the viability of outlier removal as an improvement technique are presented. In all cases, however, the results consistently show that STAPLE estimates using the robust extensions provide a consistent and accurate model of the “ground truth” from a wide range of online (Internet-based) collaborations among minimally trained human raters.

## 2. METHODS

All data presented in this paper was gathered using the WebMill interface (https://brassie.ece.jhu.edu/Home). Detailed instructions for the labeling procedure for all data sets were provided for all observations. All raters were required to perform at least one practice labeling before proceeding to the actual label sets. The implementation of STAPLE (and STAPLER) was created in Matlab (Mathwords, Natick MA). All studies were run on a 64 bit 2.5 GHz notebook with 3.2 GB of RAM. As in [5], simultaneous parameter and truth level estimation was performed with iterative expectation maximization.

### 2.1 Training and Testing Data

A testing and training set were presented to the raters for all data sets except for the pig cardiac data. Through the WebMill interface, the training and testing sets were presented slightly differently. The “ground truth” was presented to the rater after each observation in the training set so that the rater could assess their quality.

### 2.2 STAPLE Accuracy with Respect to Number of Labels

The accuracy of the STAPLE algorithm was assessed with respect to the number of labels used in the “ground truth”. This was accomplished by gathering all of the data for which a truth exists and categorizing it based upon the number of unique labels. The accuracy of each observation was assessed by calculating its percent difference with the truth.

### 2.3 Outlier Removal

Given the variability of the observations (Figure 1), the amount of improvement that could be achieved through removing outlier observations was assessed. This was accomplished by taking an omniscient perspective and keeping only observations within an arbitrary percent difference of the “ground truth”. A new STAPLE estimate was then constructed using only the observations that were found to be within this percent difference threshold.

## 3. DATA

### 3.1 Pig Cardiac Labeling

First, raters were asked to label both the epicardial and endocardial layers of a pig heart. Imaging data were acquired from a single anesthetized pig using a high resolution, steady-state free suppression (SSFP) acquisition with breath holds on a 3T Philips Achieva scanner (Best, The Netherlands). A total of 273 slices (39 volumes, 7 slices per volume, 151×151 voxels) were used for the data set. Using the WebMill system, 10 raters labeled between 150 and 500 observations. Raters ranged in experience from highly-trained experts to undergraduate imaging students. A total of 3096 observations comprised the data set. No ground truth was presented for this data set. Figure 1A shows representative observations from the pig cardiac data set. Data from all of the volumes was combined and estimations of the truth were created using STAPLER. Consistency of estimates was compared to the consistency of an arbitrary rater and expert raters.

### 3.2 Simulated Data Labeling (spheres of various radii randomly located in a cylinder)

Second, a simulated data was created to model a theoretical cylinder containing randomly placed spheres of varying radii, for example representing localization of tumors. Both a training data set and testing data set were created using the simulation data. Each data set contained 64 slices and each image was 64×64 pixels comprising a 64×64×64 3D coverage of two simulated cylinders with randomly placed spheres of varying radii. All of the raters in this experiment were minimally trained imaging students. For the training set, a total of 54 raters were used with each rater labeling between 5 and 386 slices. Each slice was labeled between 19 and 36 times (a total of 1820 training observations). For the testing set, a 45 raters each rated between 2 and 748 slices. Each slice was labeled between 77 and 52 times (a total of 4145 observations). Figure 1B shows representative observations from the simulated data label set. All of the data from all of the individual raters was combined and estimations of the truth were garnered using both traditional STAPLE and the updated STAPLER.

### 3.3 Brain Labeling (Sagittal, Axial and Coronal Cross Sections)

Finally, labeling of the cerebellum with a high resolution MPRAGE (magnetization prepared rapid acquired gradient echo) sequence was studied. Whole-brain scans of two healthy individuals (after informed written consent prior) were acquired (182×218×182 voxels), and each slice was cropped to isolate the posterior fossa. Both datasets were manually labeled by a neuroanatomical expert in a labor intensive process (approximately 20 hours each). One dataset was designated for training and one for testing. Sagittal, axial and coronal cross sections were created and presented for labeling for both data sets. Thirty-eight undergraduate students were recruited with no special knowledge of neuroanatomy. For the sagittal set, raters labeled between 0 and 119 slices (training: 583 total) and between 0 and 161 slices (1650 total). For the axial set, raters labeled between 0 and 91 slices (training: 540 total) and between 0 and 175 (testing: 1066 total). For the coronal set, raters labeled between 0 and 64 slices (training: 301 total) and between 0 and 363 slices (testing: 1532 total). Figure 1C shows example observations from these cross sections. All of the data from all of the individual raters was combined and estimations of the truth were garnered using both traditional STAPLE and the STAPLER.

## 4. RESULTS

### 4.1 Consistency Comparison - STAPLE, Experts, and Arbitrary Minimally Trained Raters

The consistency of STAPLE estimates was compared to the consistency of an expert and an additional arbitrary minimally trained rater. The consistency was found by averaging the Jaccard index [8,9] for all combinations of estimates and observations for a single slice. The results can be seen in Table 1.

### 4.2 STAPLE Accuracy using Simulation Data

Figure 2 shows that the STAPLER estimates are consistently between 98% and 99%. On the other hand, traditional STAPLE breaks down on certain slices and fails to perform better than even the lower quartile of observations.

### 4.3 STAPLE Accuracy using Sagittal, Axial and Coronal Cerebellum Data

Figures 3 and 4 show that the STAPLER estimates using the prior are consistently above the upper quartiles of the individual observations. The substantial accuracy disparity between the traditional STAPLE estimates on the training and testing set can be attributed to the spread of individual observations and the disparity in the number of observations for the training and testing sets. Similar results (not shown) were seen for both the sagittal and coronal cross sections of the image data.

### 4.4 Training Data vs. Testing Data

The results shown in Figure 3 (training data) and Figure 4 (testing data) show a significant difference in the quality of individual observations. In most cases the interquartile range on the slices in the training data is at least 3 times as large as the individual slices in the training data. This trend was also consistent with the results for both the sagittal and coronal cross sections of the brain (not shown). The reason for this phenomenon is unknown. The only difference between the training and testing data was that after an observation was made on the training data the underlying truth was shown to the rater so that they could be aware of the quality of their ratings.

### 4.5 STAPLE Accuracy with Respect to Number of Labels

As shown in Figure 5, the spread of the individual observations is substantially larger than the spread of the STAPLER estimate and the individuals observations are of generally lower accuracy. Here, we see that the number of labels can be seen as a rough proxy for the difficulty of the observation and, thus, the difficulty in creating an accurate estimate.

## 5. CONCLUSIONS

The ability to use a statistically based algorithm to accurately estimate the “ground truth” in the presence of minimally trained human foibles and follies is presented in this paper. It has now been shown that the traditional approach of using only highly trained experts to generate an accurate estimation is no longer necessary. Additionally, the importance of the STAPLER approach to estimation has been clearly demonstrated. Traditional STAPLE is dramatically susceptible to raters with a small number of observations and outlier observations. The biasing prior introduced by STAPLER stabilizes the estimations and prevents wildly inconsistent estimations.

Three types of data were presented in this paper. The pig cardiac data was used to demonstrate the dramatic consistency improvement that STAPLE provides over both an arbitrary minimally trained rater as well as an expert neuroanatomist. The simulated data clearly demonstrates the susceptibility of traditional STAPLE to outlier observations, while, at the same time, demonstrates the stabilizing quality of STAPLER’s biasing prior (consistently above 98% accuracy). Lastly, the cerebellum labeling case demonstrates the difficulty in estimating slices with large numbers of labels. Additionally, the large variation in the individual observations and the limited number of observations explain why the traditional STAPLE wildly breaks down on the training set and not on the testing set.

There are many factors that go into determining the difficulty of labeling an individual slice, however, the results indicate that as the number of labels increases, the accuracy of the estimation decreases. This is consistent with the individual observations, as the number of labels increases, the median accuracy for a set of observations decreases. While by no means a formal proof, it is consistent with the idea that the number of labels on a given slice can be potentially used as a proxy for the difficulty of delineating that slice.

Interestingly, “perfect” outlier rejection does not substantially improve the performance of STAPLER (e.g., less than 3% improvement for any of the presented experiments, data not shown). To illustrate this phenomenon, we used an omniscient perspective to remove substantially incorrect outliers (<10% correct versus the known truth). Hence, outliers do not appear to dramatically affect the reliability of accuracy of STAPLER when the using a biasing prior. This suggests that new considerations of how to parameterize human behavior may enable more efficient data use through label fusion than improvements in outlier rejection. Perhaps these experiments highlight the limitations of a traditional Gaussian model to capture the unique qualities of labeling error and randomness when presented with human foibles.

Nevertheless, it is demonstrated that the STAPLER approach as presented, which is Gaussian based, successfully performs label fusion in a variety of contexts using data from online (Internet-based) collaborations among minimally trained raters. This first successful demonstration of a STAPLE approach using “wild-type” data opens numerous possibilities for very large scale efforts in collaboration. Extension and generalization of these technologies for new application spaces will certainly present fascinating areas for continuing research.

## Acknowledgments

This project was supported by NIH/NINDS 1R01NS056307 and NIH/NINDS 1R21NS064534. The authors are grateful to Fangxu Xing, Sahar Soleimanifard, and our numerous volunteer raters for their assistance in data acquisition.

## References

1. Kearns, M.; Valiant, LG. Harvard University Technical Report, TR-14–88. 1998. Learning boolean formulae or finite automata is as hard as factoring.
2. Shapire RE. The strength of weak learnability. Machine Learning. 1990; 5(2):197–227.
3. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Computer and System Sciences. 1997; 55:119–139.
4. Warfield, SK.; Zou, KH.; Kaus, MR.; Wells, WM. Simultaneous validation of image segmentation and assessment of expert quality. Washington, DC: 2002.
5. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004; 23(7):903–21. [PubMed: 15250643]
6. Rohlfing T, Russakoff DB, Maurer CR. Expectation maximization strategies for multi-atlas multi-label segmentation. Inf Process Med Imaging. 2003; 18:210–21. [PubMed: 15344459]
7. Udupa J, LeBlanc V, Zhuge Y, Imielinska C, Schmidt H, Currie L, Hirsch B, Woodburn J. A framework for evaluating image segmentation algorithms. Comp Med Imag Graphics. 2006; 30(2): 75–87.
8. Gee JC, Reivich M, Bajcsy R. Elastically deforming 3D atlas to match anatomical brain images. J Comput Assist Tomogr. 1993; 17:225–236. [PubMed: 8454749]
9. Jaccard P. The distribution of flora in the alpine zone. New Phytol. 1912; 11:37–50.
10. Landman, BA.; Bogovic, J.; Prince, JL. Simultaneous Truth and Performance Level Estimation with Incomplete, Over-complete, and Ancillary Data.
11. Bogovic, J.; Landman, BA.; Bazin, P-L.; Prince, JL. Statistical Fusion of Surface Labels provided by Multiple Raters, Over-complete, and Ancillary Data.
12. Landman, BA.; Bogovic, J.; Price, JL. Efficient Anatomical Labeling by Statistical Recombination of Partially Label Datasets.

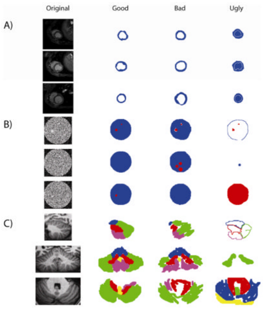

**Figure 1.**

The three major tasks presented in this paper and the gamut of their observations. The pig cardiac data can be seen in A), the simulated cylinder data can be seen in B), and the sagittal, axial, and coronal cerebellum data can be seen in C). The range of observations is broken up into three classifications. The good classification represents observations that are high quality observations given the original image slice. The bad classification represents observations where the rules were followed but the labeled images are not necessarily close to the ground truth. The ugly classification represents blatant rule breaking and observations that are completely inconsistent with the expected ground truth.

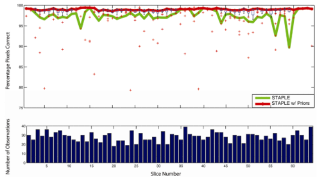

**Figure 2.**

Summary of simple simulated cylinder data. The accuracy (measured in percentage of pixels correct) of STAPLE, STAPLE using a prior, and the individual observations against the underlying ground truth can be seen in the upper row. The STAPLE estimate can be seen in green, the STAPLE estimate using the priors can be seen in red and the individual observations are represented using box plots for each slice. The total number of individual observations for each slice can be seen in the bar graph in the bottom row. Of note here is the susceptibility of traditional STAPLE to outlier observations and the high accuracy of STAPLE using priors in the simple two to three label case.

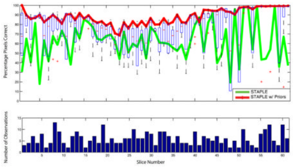

**Figure 3.**

Summary of axial cerebellum label training data. The accuracy (measured in percentage of pixels correct) of STAPLE, STAPLE using a prior, and the individual observations against the underlying ground truth can be seen in the upper row. The STAPLE estimate can be seen in green, the STAPLE estimate using the priors can be seen in red and the individual observations are represented using box plots for each slice. The total number of individual observations for each slice can be seen in the bar graph in the bottom row. The susceptibility of traditional STAPLE to outlier observations and the loss in accuracy when the difficulty of the labeling process is increased and the number of potential labels is increased from 3 (Figure 2) to 6 are the main observations here.

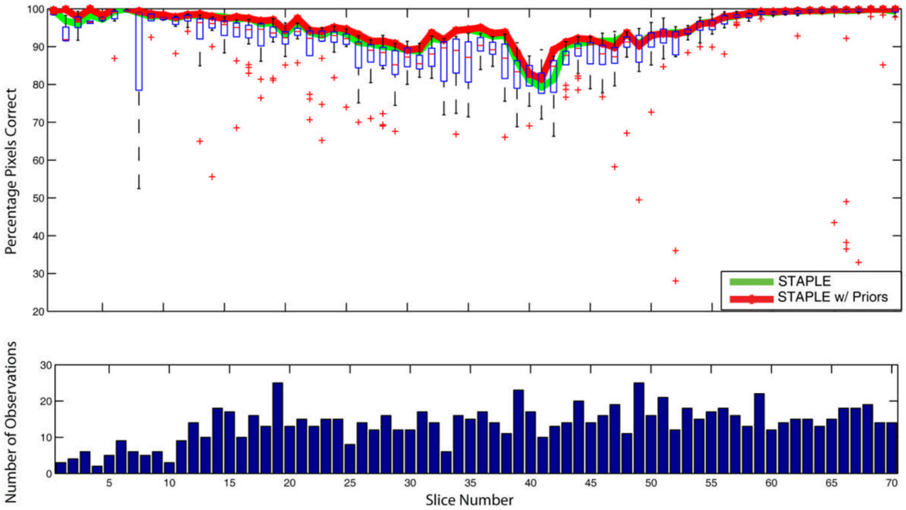

**Figure 4.**

Summary of axial cerebellum label testing data. The accuracy (measured in percentage of pixels correct) of STAPLE, STAPLE using a prior, and the individual observations against the underlying ground truth can be seen in the upper row. The STAPLE estimate can be seen in green, the STAPLE estimate using the priors can be seen in red and the individual observations are represented using box plots for each slice. The total number of individual observations for each slice can be seen in the bar graph in the bottom row. The spread of the individual observations for the testing data was significantly smaller than the spread for the training data (Figure 3) which causes traditional STAPLE to be consistent with the version of STAPLE that uses priors. Still of note is the fact that the overall accuracy of the axial cerebellum is significantly less than the accuracy of the simulation data (Figure 2).

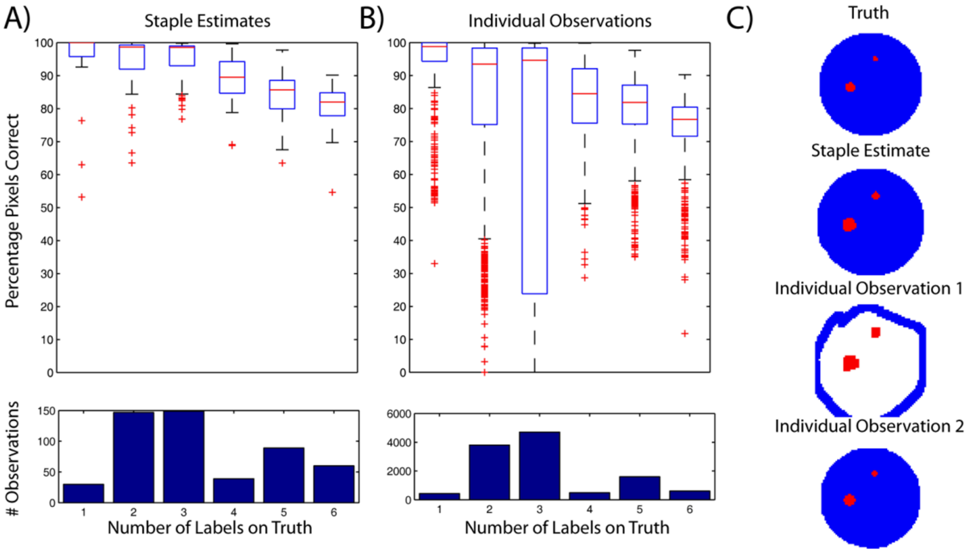

**Figure 5.**

Comparison between spread of STAPLE estimates and individual observations with respect to the number of labels present on the ground truth. The spread of STAPLE estimates (seen in A)) is represented by a box plot for each of the possible number of labels on the ground truth model. A bar graph presenting the number of estimations using this number of labels can below the box plots. The same information for the individual observations can be seen in B). Examples of a ground truth, a STAPLE estimate and two individual observations can be seen in C).

### Table 1

Comparison of consistency between arbitrary rater, expert rater and STAPLE. The consistency of each is reported as the average Jaccard index of all combinations of observations and estimates for a given slice. The uncertainty is the standard deviation of these Jaccard indices.

| | |
|---|---|
| Arbitrary Rater | 0.7377 ± 0.0525 |
| Expert Rater | 0.8003 ± 0.0480 |
| STAPLE with priors | 0.9998 ± 0.0011 |