Segmentation of malignant gliomas through remote collaboration and statistical fusion

Zhoubing Xu, Andrew J. Asman, and Eesha Singh
Electrical Engineering, Vanderbilt University, Nashville, Tennessee 37235

Lola Chambless and Reid Thompson
Neurosurgery, Vanderbilt University, Nashville, Tennessee 37235

Bennett A. Landman^{a)}
*Electrical Engineering, Vanderbilt University, Nashville, Tennessee 37235 and
Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218*

(Received 24 April 2012; revised 17 August 2012; accepted for publication 20 August 2012; published 12 September 2012)

Purpose: Malignant gliomas represent an aggressive class of central nervous system neoplasms. Correlation of interventional outcomes with tumor morphometry data necessitates 3D segmentation of tumors (typically based on magnetic resonance imaging). Expert delineation is the long-held gold standard for tumor segmentation, but is exceptionally resource intensive and subject to intrarater and inter-rater variability. Automated tumor segmentation algorithms have been demonstrated for a variety of imaging modalities and tumor phenotypes, but translation of these methods across clinical study designs is problematic given variation in image acquisition, tumor characteristics, segmentation objectives, and validation criteria. Herein, the authors demonstrate an alternative approach for high-throughput tumor segmentation using Internet-based, collaborative labeling.

Methods: In a study of 85 human raters and 98 tumor patients, raters were recruited from a general university campus population (i.e., no specific medical knowledge), given minimal training, and provided web-based tools to label MRI images based on 2D cross sections. The labeling goal was characterized as to extract the enhanced tumor cores on T1-weighted MRI and the bright abnormality on T2-weighted MRI. An experienced rater manually constructed the ground truth volumes of a randomly sampled subcohort of 48 tumor subjects (for both T1w and T2w). Raters' taskwise individual observations, as well as the volumewise truth estimates via statistical fusion method, were evaluated over the subjects having the ground truth.

Results: Individual raters were able to reliably characterize (with >0.8 dice similarity coefficient, DSC) the gadolinium-enhancing cores and extent of the edematous areas only slightly more than half of the time. Yet, human raters were efficient in terms of providing these highly variable segmentations (less than 20 s per slice). When statistical fusion was used to combine the results of seven raters per slice for all slices in the datasets, the 3D agreement of the fused results with expertly delineated segmentations was on par with the inter-rater reliability observed between experienced raters using traditional 3D tools (approximately 0.85 DSC). The cumulative time spent per tumor patient with the collaborative approach was equivalent to that with an experienced rater, but the collaborative approach could be achieved with less training time, fewer resources, and efficient parallelization.

Conclusions: Hence, collaborative labeling is a promising technique with potentially wide applicability to cost-effective manual labeling of medical images. © 2012 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4749967>]

Key words: Malignant glioma, collaborative labeling, tumor segmentation, statistical fusion

I. INTRODUCTION

Malignant gliomas are the most common form of primary neoplasm in the central nervous system, and one of the most rapidly fatal of all human malignancies. They are treated by maximal surgical resection followed by radiation and chemotherapy. The degree of surgical resection and the grade of tumor strongly correlate with recurrence rate and survival likelihood.¹ Interestingly, while evidence exists that precise tumor morphometry might lead to stronger predictive power for outcome measures,² current interventional best practices assess the preoperative tumor volume through measurement

of the RECIST criteria,³ McDonald criteria (i.e., largest tumor diameters),⁴ and postoperative degree of resection via qualitative judgment. The development of a rapid and reliable method to assess tumor volume could help standardize radiographic assessment and allow for improved comparison between outcome and efficacy studies.

Manual delineation by anatomical experts is the gold standard for tumor segmentation. Yet, it is impractical to have anatomical experts label large-scale clinical studies, especially as the benefits of characterizing tumors beyond the RECIST criteria present speculative research opportunities. Further, the MRI appearance of gliomas can be highly

variable with numerous low-contrast boundaries so that experts have been reported to suffer from intravariability and intervariability (e.g., $20\% \pm 15\%$ and $28\% \pm 12\%$ by volume, respectively⁵). Accurate, consistent, and fully automated algorithms for tumor segmentation have been the target of intense efforts for decades.^{6,7} Recent efforts have greatly increased overlap measures—near 0.9 dice similarity coefficient (DSC).^{8,9} Numerical comparisons across studies are difficult when considering heterogeneous patient populations^{8,9} and varied ranging tumor grades.¹⁰ Angelini *et al.* reviewed several approaches to tumor segmentation,¹¹ but full automation remains a largely unsettled field with reported accuracies on the order of 0.8 DSC.^{12,13}

Traditionally, manual image labeling has been achieved by individual expert raters carefully annotating each image in a series. Crowd-sourced, collaborative labeling has the potential to exploit the “wisdom of the crowd” and avoid strict requirements for particularly “wise” individuals. The recently presented WebMILL framework promises to bring together individuals (i.e., “millers”) independent of physical location and use statistical fusion to combine their results.¹⁴ Within the WebMILL framework, millers work independently on overlapping aspects of a common same segmentation project, and “collaboration” is achieved via statistical fusion of their mutual efforts. This potentially transformative technology presents a new set of challenges: investigators must pose the labeling tasks in a manner accessible to people with little or no background in medical imaging and who cannot be expected to read detailed instructions. Previously, this approach has been evaluated in toy examples¹⁵ and cerebellar labeling,¹⁴ but has not been considered for grossly abnormal anatomies (i.e., malignant brain tumors).

Herein, we characterize the efficacy of the WebMILL approach to capture two aspects of malignant glioma morphometry that are clinically relevant and well-suited for collaborative labeling. We consider volumetric segmentation of the tumor into the (1) enhancing core, which includes necrotic regions and active areas with reduced blood-brain barrier, and (2) regions of surrounding edema, which indicates inflammatory processes. These areas are characteristically visible, typically asymmetric across the brain-midline, and simple to describe a lay audience. This paper is an extension of a recent conference report.¹⁶

II. METHODS

In high grade gliomas, tumor borders are not always well-defined given diffuse invasion and infiltration into surrounding tissues. For clarity, we defined the labeling objectives to focus on clearly communicable and reproducible definitions of the extent of abnormalities. For the gadolinium-contrast enhanced T1-weighted (T1w) images, the task was to capture the area corresponding to the tumor core which is typically encompassed by active regions with reduced blood-brain barrier and signal enhancement. For the T2-weighted (T2w) images, the task was to extract the abnormally bright region (possibly diffusion) which includes the tumor core, necrosis, and surrounding edema.

II.A. Data

Clinically acquired, preoperative T1w and T2w brain MRI scans from 98 patients with high grade gliomas (WHO grade III) were acquired anonymously under Institutional Review Boards (IRB) approval. Patients had no history of prior neurosurgical intervention. On average, each T1w scan was $256 \times 256 \times 160$ voxels by the resolution of $1.0 \times 1.0 \times 1.2 \text{ mm}^3$ in the axial direction; each T2 scan was acquired axially either by resolution of $0.45 \times 0.45 \times 3.0 \text{ mm}^3$ for $512 \times 512 \times 60$ voxels, or resolution of $0.90 \times 0.90 \times 3.0 \text{ mm}^3$ for $256 \times 256 \times 60$ voxels. To enable subsequent fusion, T1w datasets were coregistered with the corresponding T2w datasets of the same patient using a rigid-body model and resampled to match the T2w resolution using the Java Image Science Toolkit¹⁷ and the Medical Image Visualization and Processing (MIPAV) platform.¹⁸

A randomly selected subset of 48 of the 98 patient datasets was selected for validation. For each dataset, an experienced graduate student manually labeled all 96 volumes (48 T1w and 48 T2w) using a high-resolution pen-input pad (Wacom, Vancouver, WA), the MIPAV three-dimensional image processing software, and a Linux workstation with high-resolution display. The graduate student was proficient with both the brain tumor anatomy and the labeling software as trained by a qualified neurosurgeon before the ground truth construction. The experienced rater spent an average of 30 min/volume (2 min 18 s/slice). To measure inter-rater variability, a subset of 10 subjects (20 volumes) was independently labeled by a second experienced rater. The second rater averaged 36 min/volume using the same computer hardware. The inter-rater reliability between the two experienced raters was 0.86 ± 0.14 DSC for T1w volumes and 0.84 ± 0.20 DSC for T2w volumes. This level of reproducibility is in agreement with previous reports which have shown that inter-rater disagreement for brain tumors is on the order of 10%–17% by volume.¹⁹

To assess accuracy of the labeled volumes from the experienced rater relative to an anatomical expert, the labels for a subset of 5 subjects (10 volumes) from the first rater were manually corrected by a neurosurgeon. The validation-after-training process is a common valid protocol for ground truth construction with minimal medical experts' involvement. The mean DSC between the labels corrected by a neurosurgeon and that of the experienced rater was 0.997 for T1w volumes and 0.9997 for T2w. Note that these exceptionally high measures reflect the bias of *correcting* truth labels rather than having an experienced rater independently create a comparative mask.

II.B. WebMILL

The WebMILL system ran as a virtualized web portal [Virtualbox (Oracle, Redwood Shores, California) with Ubuntu Linux (Canonical, Isle of Man)] in which the participants (“millers”) could access from on or off campus via a Java-enabled (Oracle, Redwood Shores, California) web browser (Fig. 1). During a 6 week period, 85 millers were recruited

FIG. 1. Flowchart of collaborative labeling approach using the WebMILL system. T1w and T2w MRI images for each patient were coregistered with a rigid-body model and resampled into the same resolution. A localizer image of a healthy volunteer at a similar slice was provided for reference. After a brief informed consent process, millers used a web-based 2D delineation tool to provide observations for the randomly selected tasks. An experienced rater and neurosurgeon generated several labeled datasets, which were shown during the training process. A statistical fusion process recombined results to form 3D volume estimates.

via flyers and e-mail advertisements. Because this research involved study of training procedures, all millers were required to undergo an in-person informed consent process. During a 10 min conversation, each miller reviewed the consent forms, created an online account, and watched a brief demonstration of the tools. After consent, millers were permitted to use the system at a self-driven pace and were compensated for up to 10 h of labeling time. At the conclusion of the study, a performance bonus was also given to the top three winners in terms of their accuracy and productivity.

Note that once WebMILL is established as a labeling process and the “work output” of the millers can be reasonably assumed, millers are likely to be considered “technical employees” and the fusion process may be considered technical evaluation of technical employees (which is generally permitted without IRB supervision). For example, one would not consent a graduate student prior to that graduate student writing a software program or soldering a circuit (although one would evaluate her work in both cases). Therefore, for *future* studies, one would likely not consider the millers as “subjects” and not need to perform informed consent for the “technical employees.”

After logging into WebMILL, millers were presented with two challenges: (1) labeling of the enhancing core on T1w MRI data and (2) labeling of the T2w enhancing region. Each challenge was structured as a continuous stream of 2D labeling tasks in which the miller would be shown a random slice of MRI data and asked to label the structure of interest. The miller could peek (but not label) up and down one slice, and a representative slice of healthy brain from the public multiparametric reproducibility study²⁰ was shown for comparison. Note that the T1 and T2 datasets were separated into two challenges because millers found task-switching difficult (i.e., labeling one T1 slice followed by a T2 slice, etc.) in early prototype studies. The user interface is setup so that a miller per-

forms training on one task, does this task for a while, and then moves onto a different task. Millers can self-pace selection of tasks between the two categories. Some millers performed more T1 sessions than T2 sessions while others decided to complete the study without performing any T2 sessions. For both sets of tasks, millers were given tasks randomly selected from the whole datasets. Tasks presented randomized slices so that facial reconstruction was not reasonably possible because no miller received all slices for all subjects and millers were not told to which subject a particular slice belonged.

Prior to entering the testing mode, millers were required to perform one (paid) practice session. In this mode, the true label would appear along with hints (i.e., outlines of the ventricles, eyes, and other anatomical features) after the user submitted their results. An additional static website provided documentation on normal brain anatomy, high grade gliomas appearance on MRI, and examples of the labeling criteria. Millers could browse these resources at any time.

III. STATISTICAL FUSION AND EVALUATION CRITERIA

We used simultaneous truth and performance level estimation with robust extensions (STAPLER) (Ref. 21) to estimate the segmentation with the collected observations from millers. As an extension of traditional simultaneous truth and performance level estimation (STAPLE) (Ref. 22) model, the algorithm is presented in an expectation maximization (EM) framework, which breaks the computation into the E-step, or the calculation of the conditional probability of the true segmentation, and the M-step, or the calculation of the rater performance parameters. In the E-step we calculate $W_{si}^{(k)}$ which represents the probability that voxel i has true label s on the k th iteration of the algorithm. In the M-step we calculate $\theta_{ss'}^{(k)}$ which represents the probability that a rater observes label s' when the true label is s on the k th iteration of the algorithm.

The STAPLER algorithm provides three basic extensions to the traditional STAPLE algorithm. These extensions are (1) the ability to take into account raters that did not observe all voxels, (2) the ability to take into account raters that observed certain voxels more than once, and (3) the ability to take into account training data (or catch-trials). As shown in Ref. 21 STAPLER was developed with the intent of utilizing data contributed by minimally trained raters, and the premise that training data is essential for estimating accurate performance level parameters.

We used DSC to validate individuals’ observations as well as our fused estimations. Herein, DSC is presented as the ratio of twice of the intersection of two segmented regions over the summation of segmented regions for both samples (slices or volumes):

$$DSC = \frac{2|A \cap B|}{|A| + |B|}, \quad (1)$$

where A and B are two segmented regions, and $|\cdot|$ represents the number of pixels/voxels of the regions. DSC is commonly considered as a standard criterion in medical image processing. Primary consideration was on slice-wise labeling of tumor

FIG. 2. The gamut of observed labels. The labeling objectives were to delineate the gadolinium enhancing tumor core (T1 images, 1st row) and label all abnormally bright tissue (T2 images, 2nd row). The experienced labeler ("truth," 2nd column) accurately found all structures, while the millers' work spanned a wide range of accuracies, which can be loosely grouped into three categories. "Good" results were qualitatively in agreement with the specified criteria. "Bad" results followed the general outline of the problem, but were loosely drawn. "Ugly" results captured gross misunderstandings.

accuracy (e.g., tumor = 1 and background = 0 for each slice) where $DSC = 1$ if both the labels indicate no tumor; however, the final evaluations consider volumetric overlap where all voxels within the imaging volume are simultaneously considered.

IV. RESULTS

For the 98 subjects, 30 615 T1w and 23 166 T2w slices were labeled in 189.4 h and 239.7 h of total system use time, respectively. For the 48 volumes which were also labeled by an experienced rater, 12 099 T1w and 8565 T2w slices were labeled. In addition to the label volumes, the system captured metrics on miller behavior for each task including the total

time spent, the number of clicks, the mouse holding time, and the time spent working with calibration and tool choice. Qualitative miller performance varied widely (Fig. 2).

IV.A. Assessment of individual millers

For the subset of 48 subjects, all miller observations were compared using the DSC on a slice-by-slice basis with the experienced rater's "truth" labels (Figs. 3 and 4). Of the 85 consented millers, 60 millers contributed T1w labels for these 48 subjects and a subset of 49 millers also contributed T2w labels. We note that millers were self-motivated and self-paced using their personal computer equipment. These rates of attrition are similar to those previously reported.

Over two-thirds of the T1w tasks and over half of the T2w tasks had a DSC over 0.80 (which we consider as "fair" performance on these datasets) (Fig. 3). The median DSC was 0.87 for T1w tasks and 0.82 for T2w tasks. We note that there were a substantial number of tasks that resulted in either perfect ($DSC = 1$) or completely incorrect ($DSC = 0$) observations; these primarily resulted from slices where the raters were shown a slice with no tumor present (so an empty response was correct while any positive response was incorrect). An examination of individual miller performances demonstrates high variance in median performance (from <0.6 to >0.9 for both tasks). Correlation between median miller performance on T1w tasks and T2w tasks was significant ($\rho = 0.5112$, $p < 0.001$). Intrarater variability was very high with nearly all raters producing some results with <0.5 DSC.

We used Fleiss' kappa²³ statistics to explain the inter-rater variations for each slice for both T1w and T2w datasets. Fleiss' kappa assesses the reliability of agreement between

FIG. 3. Overview of taskwise and millerwise accuracy. Histograms of task accuracy (a, d) illustrate generally high accuracy of miller responses based on DSC criteria. A breakdown of accuracy by rater (b, e) shows that the accuracy of specific results for individual raters was highly variable. The horizontal axis (rater #) in (b) was ordered by the median DSC on T1w tasks for each rater; this order was preserved across all plots in this paper. Fleiss' kappa (c, f) assesses the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items.

FIG. 4. Overview of taskwise and millerwise speed. Histograms for all results (a, c) show that millers typically spent less than 20 s per slice; however, a nontrivial number of tasks took more than 3 min. Investigation of speed by millers (b, d) shows a substantial spread in millers' median speed (ranging from 10 s to 40 s). As with performance, intramiller variability in speed was also large with many millers spending >90 s on tasks. In (b, d), note that the millers are ordered by T1w DSC (Fig. 3).

a fixed number of raters when assigning categorical ratings to a number of items or classifying items:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \quad (2)$$

where $\bar{P}$ is the relative observed agreement among raters, and $\bar{P}_e$ is the hypothetical probability of chance agreement. Thus $1 - \bar{P}_e$ gives the degree of agreement that is attainable by chance, and $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance.

For each slice, we identified the randomness of the decisions on each pixel from different raters. In both T1 and T2 datasets, over 60% of the slices have relatively high inter-rater agreement over 60% [Figs. 3(c) and 3(f)], while millers also gave observations with large variations on some slices.

Millers generally produced label masks very quickly per slice (Fig. 4) compared with the speed of an experienced rater. The T1w tasks had a median speed of 16 s, while the T2w tasks had a median speed of 18 s per task. Yet, approximately 1% of tasks consumed more than 3 min. These extended periods appear to be reflective of idle periods in millers work (break, distraction, etc.) as opposed to extensively careful labeling—results produced during the extended periods were not significantly better than tasks completed under 3 min (unpaired t-test, $p = 0.3624$).

Slices with larger tumor area were generally labeled more accurately [Figs. 5(a) and 5(d): T1w: $\rho = 0.2816$, $p < 0.001$, T2w: $\rho = 0.3772$, $p < 0.001$]. Slices with larger tumor areas took more time to label [Figs. 5(b) and 5(e): T1w: $\rho = 0.4395$, $p < 0.001$, T2w: $\rho = 0.3969$, $p < 0.001$]. The median time that raters spent per slice was significantly correlated with the median performance [Figs. 5(c) and 5(f): T1w: $\rho = 0.3048$, $p = 0.0179$, T2w: $\rho = 0.3158$, $p = 0.0271$]. The strength of time-DSC correlation was greater when considering results with DSC over 0.90 [Figs. 5(c) and 5(f) Inlay: T1w: $\rho = 0.6090$, $p < 0.001$, T2w: $\rho = 0.6202$, $p < 0.001$].

IV.B. Assessment of collaborative efforts

We evaluated the accuracy of the WebMILL system as a function of the number of times each slice is labeled by millers ("coverages"). One coverage is when all slices are labeled exactly once, two coverages is when all slices are labeled exactly twice, etc. The labels for slices from multiple raters were fused with the STAPLER (Ref. 21) algorithm (run for one estimation/maximization iteration) in which the probability of an error by each rater is estimated by pooling data across slices and then rater decisions are combined based upon a weighted vote scaled by rater reliability.²¹ The

FIG. 5. Accuracy–speed correlations. Except for reports of empty slices, slices with larger tumor size had higher overall DSC and were easier to label across millers (a, d). Slices with larger tumors took relatively more time to label (b, e). In general, millers who spent more time on a task were more likely to provide more accurate observations (c, f).

complete data set consisted of an average of 38 T1w coverages and 27 T2w coverages.

We used Monte Carlo resampling of all data (without replacement with a dataset subset) to construct ten unique datasets with between one and nine coverages (90 datasets total) and fused the results. In other words, a single coverage constructed by random selection of labeled observations was resampled in a slice basis. Median DSC accuracy evaluated over tumor volumes sharply increased from near 0.8 (one coverage = no redundant information per slice) to 0.9 (nine coverages = each pixel labeled nine times) for both T1 and T2 tasks (Fig. 6). Note that one coverage is approximately equal to $1/8^{\text{th}}$ the time spent by an experienced rater while nine coverages is approximately 15% longer than a single experienced rater. Fusion of at least five coverages yields inter-rater reliabilities accuracies on par with those reported in the literature

for experienced raters [coefficient of variation approximately 15% (Ref. 24)]. To complement interpretation of DSC measures, surface distances were calculated between the labeled volumes drawn by the experience rater and the estimated volume (Fig. 7).

V. LIMITATIONS AND CONSIDERATIONS

We note that 5 of 48 volumes for both T1 and T2 exhibited poor DSC per tumor volume even with nine coverages (labeled a–e in Fig. 6). Examination of these particular “modes of failure” of the WebMILL system is revealing. Note that we show representative slices from *every* outlier case so that the reader can appreciate the modes of failure.

FIG. 6. Mean accuracy for 48 patients of fused miller estimates (over 10 Monte Carlo trials) increases for the 48 subject brains included in each of the T1 (a) and T2 (c) labeling tasks. Representative renderings of fused labels (light) and truth (dark) for the tumor core (b) and T2 abnormality (d) show that the 3D shape largely agrees without visually apparent bias. The outliers with low DSC measures (points labeled a–e) are explored in Fig. 8.

FIG. 7. Surface distances evaluation. Surface distances were calculated between the ground truth and the estimated volume, which demonstrated a mean improvement in surface distances of approximately 10% for seven coverages over one coverage.

For the T1 outliers (Fig. 8, 1st row):

The gadolinium enhancements in (a), (b), and (e) are very minor. The majority of raters outlined the vacuoles in (a), but did not include the non-enhancing core. In (b), it is up to interpretation as to whether the region labeled by the expert exhibits the enhancing core criteria. In (e), the combined miller result more clearly encompasses the enhancing region.

Slice (c) shows an excellent example where one could easily mistake the tumor core necrosis for a ventricle on first glance. The majority of raters pressed “next” and did not label any pixels on this slice. We note that a substantial minority of millers correctly labeled this region.

Slice (d) shows an unusual manifestation in which the millers identified the tumor core but omitted the surrounding region where the experienced labeler included abnormal ventricle tissues and areas of nonenhancement.

For the T2 outliers (Fig. 8, 2nd row)

Slice (a) illustrates the distinction between abnormal enhancement and normal fluid can be challenging when the ventricles are involved. Slice (d) presents a similar case in which enhanced gray matter is highly similar to normal fluid.

Slices (b) and (c) show relatively large tumors with diffuse boundaries. In these cases, the millers tended to be more conservative in declaring abnormal tissue.

Slice (e) illustrates a case in which millers focused on the obvious necrotic vacuole but did not seek to include additional areas of enhancement.

These modes of failure highlight the weakness of this approach to capture outlier cases.

VI. DISCUSSION

In summary, we demonstrate that the collaborative efforts of minimally trained raters can identify the core and edema regions of malignant gliomas with equivalent accuracy to that of an experienced graduate student and staff rater (≈ 0.85 DSC), as well as in terms of intervariability and intravariability. The collaborative approach is highly parallel and efficient in terms of time (i.e., the total time spent by the collective is equivalent to that of a single expert) and resources (i.e., only minimal training and no hardware is provided to the participants). Hence, collaborative labeling is a very promising new technique with potentially wide applicability to cost-effective manual labeling of medical imaging data as defined in the following paragraphs.

Careful analysis of “failure” outliers highlights important characteristics of collaborative labeling: (1) Millers do not readily incorporate external expectations of tumor characteristics into the labeling process and tended to differ from the experienced rater when the presence of enhancing features was highly subjective. (2) An empty result provided by a miller should be interpreted with less confidence than a region that has been carefully labeled as millers completely miss subtle results more frequently than they would misplace a label.

FIG. 8. Visual comparison between truth (dark) and typical observations from ten fused millers (light) for results with low DSC (see Sec. II.B).

(3) Conflicting regions for enhancement (i.e., normal fluid versus T2 abnormality) can cause substantial confusion and it would likely be worthwhile to account for multiparametric tissue contrast during the fusion process. (4) As with experienced raters, millers are human and sometimes simply make mistakes; consideration of 3D consistency in the fusion process would likely improve results. These observations present opportunities for further optimization of the statistical fusion model and improved utilization of collaborative labeling efforts. Except for the demonstrated outlier cases (5/48), fusion of millers datasets provide reasonably good segmentations.

Labeling with minimally trained raters can be problematic: (1) millers might not follow instructions, (2) millers might not *want* to follow instructions, and (3) millers might bias the distribution by providing out-of-range results. In fact, we observed all three aspects of these behaviors as illustrated in Fig. 2 for the “ugly” results. Fortunately, and perhaps surprisingly, the statistical fusion approach was robust to these outliers. We did not remove any of the “ugly” results from analysis and did not provide the fusion algorithm any exogenous information as to outlier/erroneous results. Nevertheless, the “ugly” observations did not substantively impact the fused results as the fusion methods automatically estimate and deweight abnormal observations.

While the STAPLER statistical fusion approach proved effective at combining information across millers, continued innovation in statistical fusion algorithms may lead to more efficient fusion. In particular, we note the possibility for enhancing 3D consistency and optimizing selection of catch trials. To date, no fusion algorithm can take into account ancillary metrics of rater behavior (time spent labeling, time spent idle, number of clicks, etc.). These metrics are simultaneously recorded for all WebMILL results and may prove effective for improving performance estimates at a level of individual tasks (as opposed to millers). In this effort, intramiller variance and variation in task “difficulty” overwhelmed the slice-by-slice predictive power using standard correlation analysis. Nevertheless, it is clear that some millers are better and some tasks are easier (Figs. 3–5); significant correlation has been found between DSC and speed for all tasks (T1w: $\rho = 0.0223$, $p = 0.0142$, T2w: $\rho = 0.0813$, $p < 0.001$). It is an open and interesting question as to how to best incorporate these findings in the statistical fusion process.

We do not anticipate that the demonstrated collaborative labeling process will have immediate applicability in clinical care given the stochastic nature of the labeling process and uncertainties in rater performance. Rather, we see WebMILL as a novel practical perspective for looking at difficult clinical problems (e.g., high grade brain tumor segmentation) with large collections of datasets which would be impractical (i.e., “burn out” experience labelers) or impossible (i.e., insufficient resources to accomplish the tasks given lab hardware) using traditional techniques. The conclusion of this work is that combined efforts of millers can be made as reliable as an experienced rater (as opposed to an anatomist expert), so we believe that the presented comparison is justified.

In practice, we have found that manual segmentations are often drawn by research assistants and reviewed by an

anatomist. The degree of anatomist involvement has widely varied across the studies with which we have been involved. Hence, WebMILL could be practically integrated into the clinical workflow as an initial labeling step; experienced labelers and/or anatomists could identify and correct the (infrequent) outliers and still achieve a net savings. Additional consideration of hybrid strategies and multistage quality assurance is warranted.

In addition to direct segmentation, we envision several promising avenues for application of collaborative labeling. (1) Large scale pilot data could be created through collaborative labeling and used to train fully automated techniques ranging from multiatlas labeling to active shape modeling. (2) Collaborative labeling could be performed by professional staff (instead of millers with limited medical background), which could allow for remote effort, intrinsic quality assurance, and potential resource-savings over hiring a single expert anatomist. (3) Collaborative labeling could be applied to correct/perform preprocessing procedures that are not ideally solved by automatic procedures (but are straightforward to human beings)—e.g., brain skull stripping.

ACKNOWLEDGMENTS

The authors appreciate the assistance of Wade Allen. This work was supported, in part, by NIH/NINDS 1R21NS064534, NIH 2R01EB006136, and NIH R01EB006193. The authors are grateful to the anonymous reviewers for their insightful comments and assistance in improving this paper.

^{a)} Author to whom correspondence should be addressed. Electronic mail: bennett.landman@vanderbilt.edu

¹ R. Sawaya, “Extent of resection in malignant gliomas: a critical summary,” *J. Neuro-Oncol.* **42**(3), 303–305 (1999).

² S. T. Astner *et al.*, “Tumor shrinkage assessed by volumetric MRI in the long-term follow-up after stereotactic radiotherapy of meningiomas,” *Strahlenther. Onkol.* **186**(8), 423–429 (2010).

³ P. Therasse, E. A. Eisenhauer, and J. Verweij, “RECIST revisited: A review of validation studies on tumour assessment,” *Eur. J. Cancer* **42**(8), 1031–1039 (2006).

⁴ W. I. McDonald *et al.*, “Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis,” *Ann. Neurol.* **50**(1), 121–127 (2001).

⁵ G. P. Mazzara *et al.*, “Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation,” *Int. J. Radiat. Oncol., Biol., Phys.* **59**(1), 300–312 (2004).

⁶ M. Vaidyanathan *et al.*, “Comparison of supervised MRI segmentation methods for tumor volume determination during therapy,” *Magn. Reson. Imaging* **13**(5), 719–728 (1995).

⁷ P. Gibbs *et al.*, “Tumour volume determination from MR images by morphological segmentation,” *Phys. Med. Biol.* **41**(11), 2437–2446 (1996).

⁸ V. Harati, R. Khayati, and A. Farzan, “Fully automated tumor segmentation based on improved fuzzy connectedness algorithm in brain MR images,” *Comput. Biol. Med.* **41**(7), 483–492 (2011).

⁹ A. Hamamci *et al.*, “Cellular automata segmentation of brain tumors on post contrast MR images,” *Med. Image Comput. Comput. Assist. Interv.* **13**(Pt 3), 137–146 (2010).

¹⁰ K. E. Emblem *et al.*, “Automatic glioma characterization from dynamic susceptibility contrast imaging: Brain tumor segmentation using knowledge-based fuzzy clustering,” *J. Magn. Reson. Imaging* **30**(1), 1–10 (2009).

- ¹¹E. Angelini *et al.*, “Glioma dynamics and computational models: A review of segmentation, registration, and in silico growth algorithms and their clinical applications,” *Curr. Med. Imaging Rev.* **3**(4), 262–276 (2007).
- ¹²J. Egger *et al.*, Nugget-cut: A segmentation scheme for spherically- and elliptically-shaped 3D objects, in *Proceedings of the 32nd DAGM Conference on Pattern Recognition* (Springer-Verlag, Darmstadt, 2010), pp. 373–382.
- ¹³T. M. Hsieh *et al.*, “Automatic segmentation of meningioma from non-contrasted brain MRI integrating fuzzy clustering and region growing,” *BMC Med. Inf. Decis. Making* **11**, 54 (2011).
- ¹⁴B. A. Landman *et al.*, “Foibles, follies, and fusion: Web-based collaboration for medical image labeling,” *Neuroimage* **59**(1), 530–539 (2011).
- ¹⁵J. A. Stein, A. J. Asman, and B. A. Landman, “Characterizing and optimizing rater performance for Internet-based collaborative labeling,” *Proc. Soc. Photo-Opt. Instrum. Eng.* **7966**, 79660M (2011).
- ¹⁶Z. Xu *et al.*, “Collaborative labeling of malignant glioma,” in *Proceedings of the 2012 International Symposium on Biomedical Imaging (ISBI)*, Barcelona, Spain, 2012, pp. 1148–1151.
- ¹⁷B. C. Lucas *et al.*, “The Java Image Science Toolkit (JIST) for rapid prototyping and publishing of neuroimaging software,” *Neuroinformatics* **8**(1), 5–17 (2010).
- ¹⁸M. J. McAuliffe *et al.*, “Medical image processing, analysis and visualization in clinical research,” in *Proceedings of the 14th IEEE Symposium on Computer-Based Medical Systems*, Bethesda, MD, 2001, pp. 381–386.
- ¹⁹B. N. Joe *et al.*, “Brain tumor volume measurement: comparison of manual and semiautomated methods,” *Radiology* **212**(3), 811–816 (1999).
- ²⁰B. Landman *et al.*, “Multi-parametric neuroimaging reproducibility: A 3T resource study,” *Neuroimage* **4**(14), 2854–2866 (2011).
- ²¹B. Landman, *et al.*, “Robust Statistical Fusion of Image Labels,” *IEEE Trans. Med. Imaging* **31**(2), 512–522 (2011).
- ²²S. 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).
- ²³J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychol. Bull.* **76**(5), 378–382 (1971).
- ²⁴M. R. Kaus *et al.*, “Automated segmentation of MR images of brain tumors 1,” *Radiology* **218**(2), 586–591 (2001).