COLLABORATIVE LABELING OF MALIGNANT GLIOMA

Zhoubing Xu[1], Andrew J. Asman[1], Eesha Singh[1],
Lola Chambless[2], Reid Thompson[2], Bennett A. Landman[1,3]

[1]Electrical Engineering, Vanderbilt University, Nashville, TN, 37235
[2]Neurosurgery, Vanderbilt University, Nashville, TN, USA 37235
[3]Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218

ABSTRACT

Malignant gliomas represent an aggressive class of central nervous system neoplasms which are often treated by maximal surgical resection. Herein, we seek to improve the methods available to quantify the extent of tumors as seen on magnetic resonance imaging using Internet-based, collaborative labeling. In a study of clinically acquired images, we demonstrate that teams of minimally trained human raters are able to reliably characterize the gadolinium-enhancing core and edema tumor regions (Dice ≈ 0.9). The collaborative approach is highly parallel and efficient in terms of time (the total time spent by the collective is equivalent to that of a single expert) and resources (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 facilitate cost-effective manual labeling of medical imaging data.

Index Terms— Collaborative Labeling, Statistical Fusion, Malignant Glioma, Segmentation

1. INTRODUCTION

Traditionally, medical image labeling has entailed individual expert raters carefully annotating each image in serial (one after another), which is resource intensive. 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 [1]. 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 examples and cerebellar labeling, but has not been considered for grossly abnormal anatomies (i.e., malignant brain tumors).

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 [2]. Interestingly, while there is evidence that precise tumor morphometry might lead to stronger predictive power for outcome measures, e.g., [3], current interventional best practices assess the pre-operative tumor volume through measurement of the RECIST criteria [4], McDonald criteria (i.e., largest tumor diameters) , and post-operative degree of resection via qualitative judgment. The development of a rapid and reliable method to assess tumor volume might help standardize radiographic assessment and allow for improved comparison between outcome and efficacy studies. Given the difficulties of automation, and the divergent accuracy metrics [5], manual or semi-automated voxel-wise labeling is typically employed to access tumor characteristics.

Herein, we characterize the efficacy of the WebMill approach to capture two aspects of malignant glioma morphometry that are both clinically relevant and well-suited for collaborative labeling. First, the core of a tumor is the minimal possible definition of tumor extent and includes the most necrotic tissue. It often appears with large vacuoles and diminished blood-brain barrier. On gadolinium-enhanced T1-weighted MRI, this region is stereotypically surrounded by a bright rim indicative of gadolinium accumulation. Second, we characterize the maximal extent of tumor growth that can be captured by the region of bright abnormality on T2-weighted MRI, which includes both tumor infiltration and surrounding edema. These attributes are characteristically visible, typically asymmetric across the brain-midline, and simple to describe to a lay audience.

2. METHODS AND RESULTS

2.1. Data

Pre-operative T1- weighted and T2-weighted brain MRI scans from 48 patients with malignant gliomas were

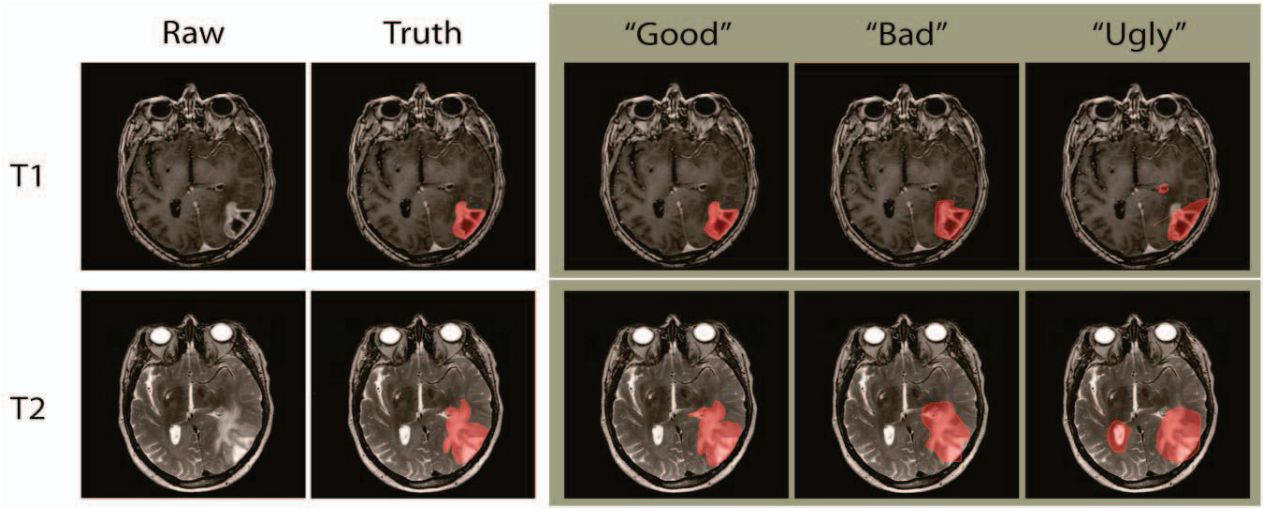


Fig 1. The gamut of observed slices. The labeling objective was to a) delineate the gadolinium enhancing tumor core (T1 images, 1st row) and b) label all abnormally bright tissue (T2 images, 2nd row). The experienced labeler ("truth", 2nd column) accurately found all structures, while the millers' work spanned a wide range of accuracies.

acquired in anonymous form under Institutional Review Boards (IRB) approval. The T1 images were co-registered to the T2 images with rigid-body registration and resampled to match resolutions. In T1 images, we defined the labeling objective as to capture any gadolinium enhancing regions (if present) corresponding to the tumor core, while, in T2 images, we defined the task to label the extent of the abnormally bright region (including both tumor and edema).

To provide ground-truth labels, an experienced graduate student manually labeled all datasets using an image-processing workstation, high resolution pen-input pad and the 3-D MIPAV software package (http://mipav.cit.nih.gov/). The experienced rater spent an average of 30 minutes per volume (mean 13 slices with 2 min 18s per slice). There were 638 slices total per T1 and T2 task.

To verify the *validity* of ground-truth labels, an experienced neurosurgeon manually corrected labels for 5 T1 and T2 volumes each. The mean Dice Similarity Coefficient (DSC) [6] over T1 brains was 0.997 and over T2 was 0.9997. Note that these exceptionally high measures are of the accuracy of the truth labels rather than what another experienced rater would have said without guidance. Inter-rater agreement for tumors has been shown to be on the order of 10-17% by volume [7].

2.2. Collaborative Labeling

Eighty-five Vanderbilt community members were recruited via posted advertisements to participate as millers over a six week period. Participation was limited to adult community members with access to the Internet and a Java compatible web-browser. Millers were paid hourly up to 10 hours and a performance bonus was paid to the top three millers in terms of total correctly labeled voxels. In an initial ten minute face-to-face conversation, millers provided informed, written consent (under IRB approval), were shown how to log into the WebMILL website, and how to use the 2-D labeling tool where previous and next slices are accessible. Participants were given a basic introduction on how to use the system, what to label, and where to find additional instructions. Millers were required to open a practice labeling session complete with annotated correct answer and perform at least one trial before proceeding to the challenge datasets. Millers were asked to self-pace and, if possible, evenly split their time between labeling T1 and T2 datasets.

Of the 85 consented participants, 60 millers (70% retention) contributed 12,099 slices for the slices corresponding to the 48 T1 weighted volumes in a total time of 87 hours 17 minutes (median time per slice of 16s). For the T2 task, millers contributed 8,565 slices in a total time of 68 hours 30 minutes (median time per slice of 18s). Some millers produced exceptionally accurate results while other produced loose boundaries or labeled structures that would clearly be considered normal by an individual experienced with the data (Fig. 1).

Fig. 2 summarizes performance by rater and performance by slice on an individual basis. Several of the millers were exceptionally good (DSC>0.9 on average), but most raters showed mean average DSC near 0.75. The average DSC is strongly influenced by a large number of tasks for which the millers did not catch the tumor (hence a DSC of 0). Average DSC for slices where raters did not miss the tumor was closer to 0.9. A linear correlation of time and performance was not observed.

We evaluated the accuracy of the WebMILL system as a function of the number of times each slice is labeled by a miller ("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 algorithm (run for one estimation/maximization iteration) in which the

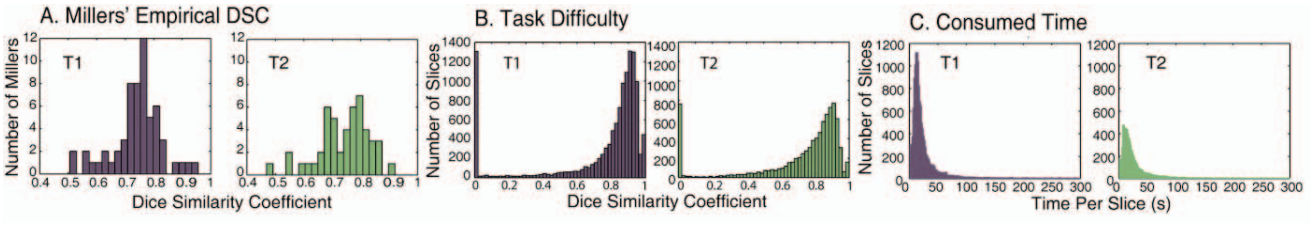


Fig 2. Summary performance measures. (A) demonstrates that some raters were exceptional while others poor, while summary measures per slice, (B), showed that some slices were more difficult than others across raters. Examinations of time spent (independent of rater) show a unimodal distribution.

probability of an error by each rater is estimated by pooling data across slices and then rater decision are combined based upon a weighted vote scaled by rater reliability [8]. The complete data set consisted of an average of 38 T1 coverages and 27 T2 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. 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. 3). Note that one coverage is approximately equal to 1/8th the time spent by an experienced rater while nine coverages is approximate 15% longer than a single rater. Fusion of at least five coverages yields inter-rater reliabilities accuracies on par with those reported in the literature for experienced raters (CV around 15% [9]).

2.2. Limitations and Considerations

We note that 5 of 48 volumes for both T1 and T2 exhibited poor DSC per tumor volume even with ten coverages (labeled A->E in Fig. 3AC). Examination of these particular "modes of failure" of the WebMILL system is revealing.

For the T1 outliers (Fig. 4, 1st row):

i. 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.

ii. 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.

iii. 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 non-enhancement.

For the T2 outliers (Fig. 4, 2nd row)

iv. A illustrates the distinction between abnormal enhancement and normal fluid can be challenging when the ventricles are involved. D presents a similar case in which enhanced grey matter is highly similar to normal fluid.

v. B and C show relatively large tumors with diffuse boundaries. In these cases, the millers tended to be more conservative in declaring abnormal tissue.

vi. 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.

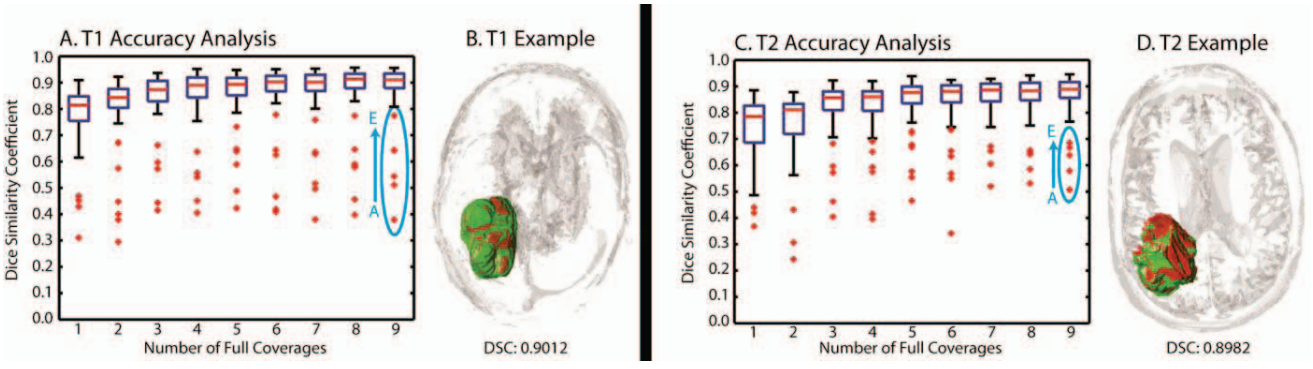


Fig 3. Mean accuracy for 48 patients of fused miller estimates (over 10 Monte Carlo trials) increases for the 48 subject brains included in each of the T1 (A) and T2 (C) labeling tasks. Representative renderings of fused labels (green) and truth (red) for the tumor core (B) and T2 abnormality (D) show that the 3-D shapes largely agree without visually apparent bias. The outliers with low DSC measures (points labeled A->E) are explored in Fig. 4.

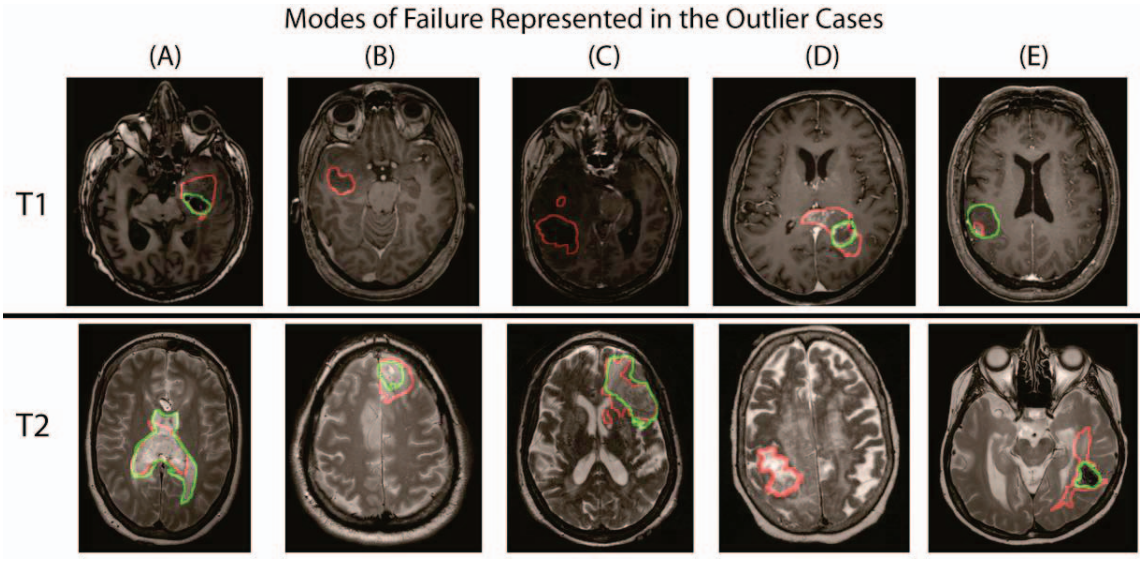


Fig 4. Visual comparison between truth (red) and typical observations from 10 fused millers (green) for results with low DSC (see section 2.2).

3. DISCUSSION

WebMILL produces volumetric labels of enhancing tumor cores and T2 signal abnormalities with promising accuracy (DSC around 0.9). The collaborative approach is highly parallel and efficient in terms of time (the total time spent by the collective is equivalent to that of a single expert) and resources (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 facilitate cost-effective manual labeling of medical imaging data. Detailed comparison with inter-rater variability for experienced on the same protocols is warranted and on-going.

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 (2.2.i). (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 (2.2.ii-iii). (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 multi-parametric tissue contrast during the fusion process (2.2.iv-v). (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 (2.2.iii,vi). These observations present opportunities for further optimization of the statistical fusion model and improved utilization of collaborative labeling efforts.

ACKNOWLEDGEMENTS: This work was supported in part by NIH/NINDS 1R01NS056307 and NIH/NINDS 1R21NS064534.

4. REFERENCES

[1] B. A. Landman, A. J. Asman, A. G. Scoggins, J. A. Bogovic, J. A. Stein, and J. L. Prince, "Foibles, follies, and fusion: Web-based collaboration for medical image labeling," *NeuroImage,* vol. 59, pp. 530-9, Jan 2011.

[2] R. Sawaya, "Extent of resection in malignant gliomas: a critical summary," *J Neurooncol,* vol. 42, pp. 303-5, May 1999.

[3] S. T. Astner, M. Theodorou, M. Dobrei-Ciuchendea, F. Auer, C. Kopp, M. Molls, and A. L. Grosu, "Tumor shrinkage assessed by volumetric MRI in the long-term follow-up after stereotactic radiotherapy of meningiomas," *Strahlenther Onkol,* vol. 186, pp. 423-9, Aug 2010.

[4] P. Therasse, E. A. Eisenhauer, and J. Verweij, "RECIST revisited: a review of validation studies on tumour assessment," *Eur J Cancer,* vol. 42, pp. 1031-9, May 2006.

[5] J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated bayesian model classification," *Medical Imaging, IEEE Transactions on,* vol. 27, pp. 629-640, 2008.

[6] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. G. Hill, and D. J. hawkes, "A comparison of similarity measures for use in 2-D-3-D medical image registration," *Medical Imaging, IEEE Transactions on,* vol. 17, pp. 586-595, 1998.

[7] B. N. Joe, M. B. Fukui, C. C. Meltzer, Q. S. Huang, R. S. Day, P. J. Greer, and M. E. Bozik, "Brain tumor volume measurement: comparison of manual and semiautomated methods," *Radiology,* vol. 212, pp. 811-6, Sep 1999.

[8] B. Landman, A. Asman, A. Scoggins, J. Bogovic, F. Xing, and J. Prince, "Robust Statistical Fusion of Image Labels," *IEEE TMI,* vol. 31, pp. 512-22, 2011.

[9] M. R. Kaus, S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, and R. Kikinis, "Automated Segmentation of MR Images of Brain Tumors," *Radiology,* vol. 218, p. 586, 2001.