AbstractID: 13515 Title: Segmentation evaluation in the context of inter-physician variance

Purpose: To evaluate a registration-based system for intracranial segmentation in the context of multiple physicians and cases of large space-occupying brain lesions. **Method and Materials:** Segmentation analyses suffer from the absence of a known ground truth and a single adequate comparison metric. However, these limitations can be overcome by applying multiple metrics and statistical methods to large data sets. We enlisted 8 physicians to delineate the brainstem, chiasm, optic nerves, and eyes in 20 challenging patient volumes. Our system utilizing the adaptive bases algorithm automatically segmented these structures. The expert delineations formed probability maps (p-maps) that were used to calculate a simulated ground truth. We evaluated inter-physician and automatic-physician variation using the Dice similarity coefficient. Each was then evaluated against the ground truth using Dice and Euclidean distance maps. To understand the observed variance at the edges of delineations, we sampled the p-maps and fitted a linear regression model. **Results:** Across all structures and patients in pair-wise Dice comparisons, the automatically derived structures compared as well or better to the physician delineations than did the physicians with one another. The inter-physician spatial overlap for brainstem and eyes was over 80%, while for the optic chiasm and nerves it was only 40-50 percent. Median Euclidean distances to a simulated ground truth were less than 2 mm for both the automatic and physician delineations, and maximum deviations were generally less than 5 mm. Of the total p-map variance only 20% was explained by main effects of physician, case and structure, and their interactions, suggesting the delineations are inherently noisy. **Conclusion:** The fully automatic system produced segmentations geometrically equivalent to those delineated by a group of physician experts. This work underscores the need for multiple physicians and metrics to avoid bias in evaluating segmentation results.