Acceptable values of similarity coefficients in neuroanatomical labeling in MRI

Andrew Worth[1], Jason Tourville[2]

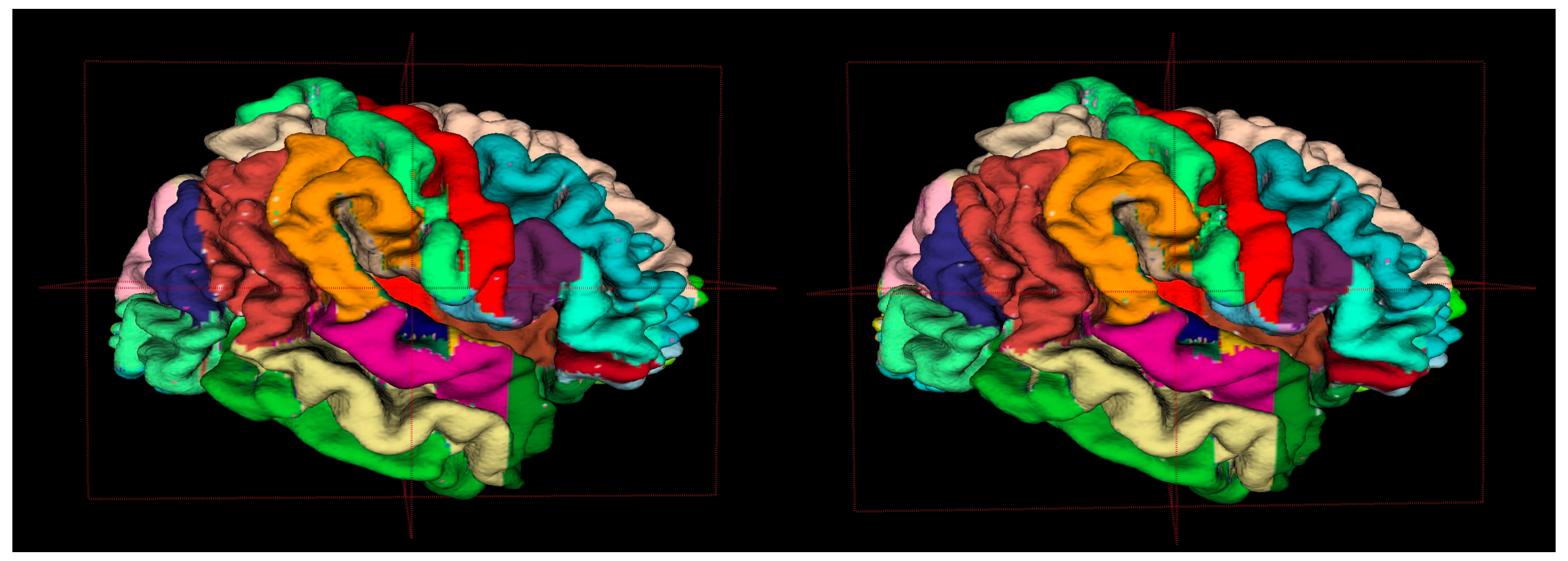
[1]Neuromorphometrics, Inc., Somerville, MA, [2]Dept. of Speech, Language, and Hearing Sciences, Boston University, Boston, MA

Introduction

There are many metrics for evaluating 3D medical image segmentation [1]. We present Dice Similarity Coefficients (DSCs) obtained from manually labeled repeat scans of 20 subjects. Each subject was scanned twice separated by some time and both scans were labeled independently and then corrected.

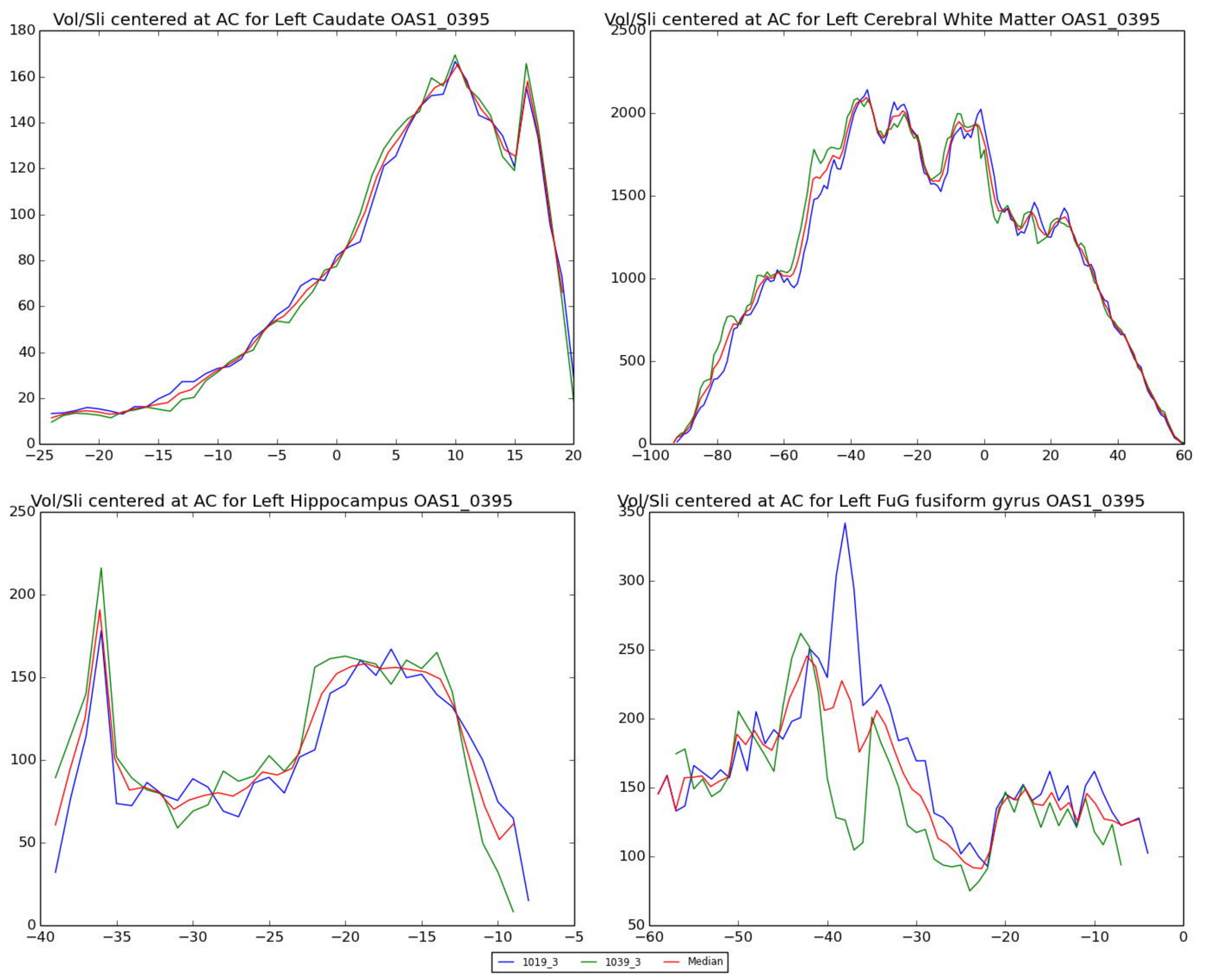
Methods

Human MRI brain scans were obtained from the "reliability" set of the Open Access Series of Imaging Studies (OASIS). We compared slice-by-slice volume plots relative to the anterior commisure for each structure looked at 3D renderings of the two scans for the same subject to flush out labeling errors. There was no attempt was to make every single line exactly the same or make pairs of borders as smooth as each other. The similarity coefficients we present thus represent a kind of upper bound on what is currently attainable. ANTs was used to register same-subject MRI scans, then the same transformation was used to register the labels [2]. Overlap statistics were calculated using ITK as described in [3].



Results

DSCs are presented for subcortical regions and also for cortical parcellations and are sorted by their median values for each structure. Similarity values occur over a range because specific anatomical regions have differing amounts of anatomical variation and are affected differently by scanning artifacts: some regions are harder to label than others because the boundaries are less apparent.



Discussion

The main advantage of manual labeling is that anything that can be seen by the human visual system can be labeled. But it is tedious, requires a lot of expertise, and can have errors due to fatigue and variation in the understanding or application of the labeling protocol. It is extremely difficult to hand-draw a boundary consistently. The main advantage of automation is that it is easy, inexpensive, and does not require the same expertise. The disadvantage is that it needs to be checked and corrected, and fails on unfamiliar anatomy.

Variation reflected in overlap metrics arises from many sources. We found pairs of scans for the same subject that were a little different, particularly in inferior regions. Two general types of intrinsic anatomical variability were observed: 1) differences in size, shape, and precise border locations, and 2) categorically different anatomical configurations. An example is the cingulate, which may appear as a single or double gyrus. Here, boundary placement can be ambiguous even with extensive manual attention.

When comparing manual and automated results, both have to have been "trained" on the exact same labeling protocol. A person with sufficient expertise can recognize atypical anatomy while most automation has no indication that anything might be wrong.

The correct "understanding" of anatomy by humans and automation can be very difficult. Because of this, the most important part of our correction procedure was to arrive at the same anatomical interpretations in both same-subject scans. However, forcing the two labeled scans to match each other does not mean the interpretation is true. The expert may have a fallacy about the exact anatomy that carries over to both.

Finally, does the extra effort matter? If the anatomy is more "correct" will a subsequent volumetric or functional analysis be significantly different and improved? Overlap and repeatability measurements in general are less important than the correlation of measurements with some final outcome. However, better test–retest reliability does suggest better precision of single measurements.

Conclusion

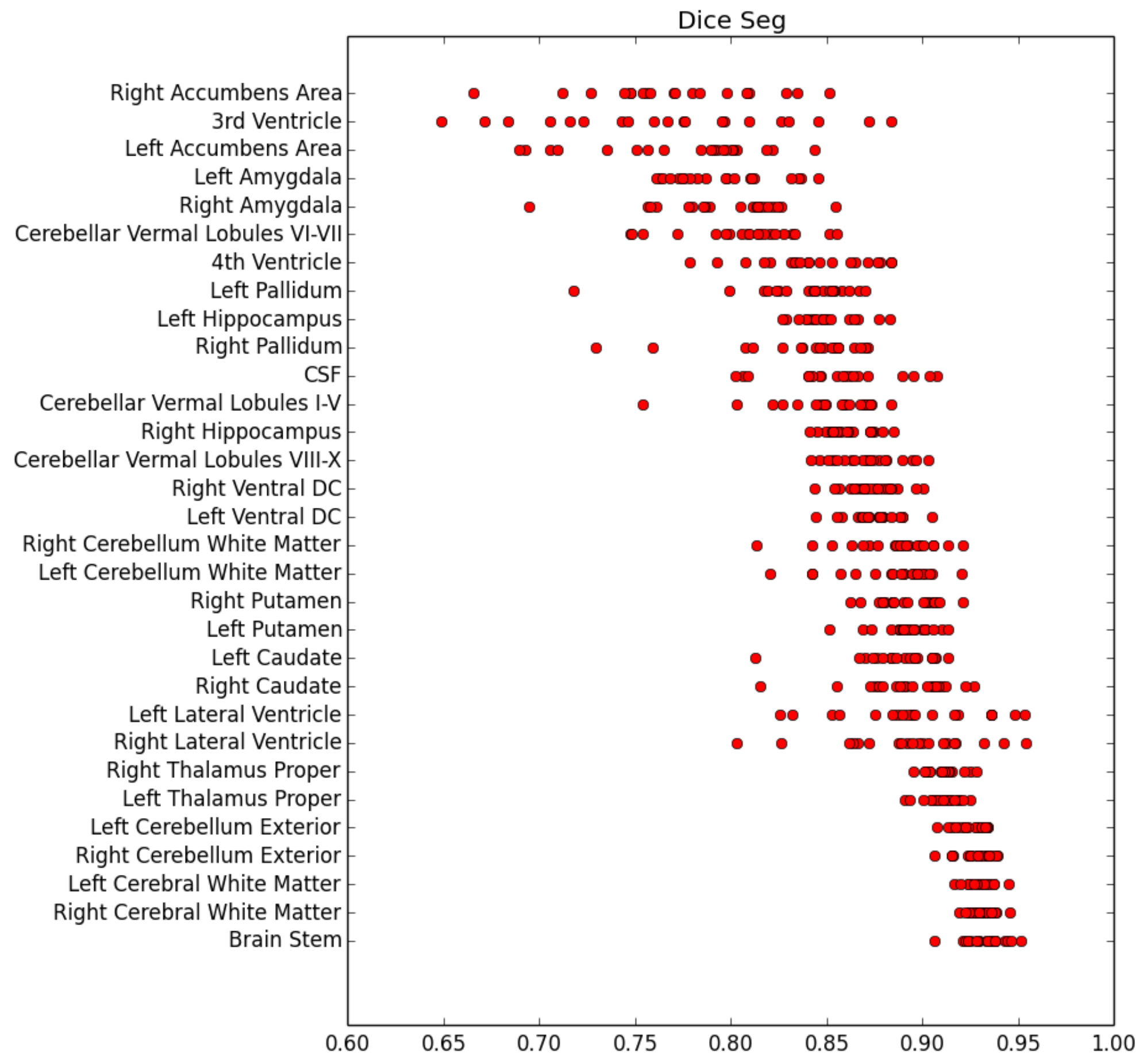
Our experience suggests that neuroanatomical labeling systems can be optimized by combining algorithms with manual checks and corrections: humans are better at interpreting anatomy and automation is better at delineating anatomy by applying that interpretation. A combination of the strengths of manual and automated methods will lead to improved labeling results, and the amount of automation will increase by algorithmically codifying additional anatomical knowledge.

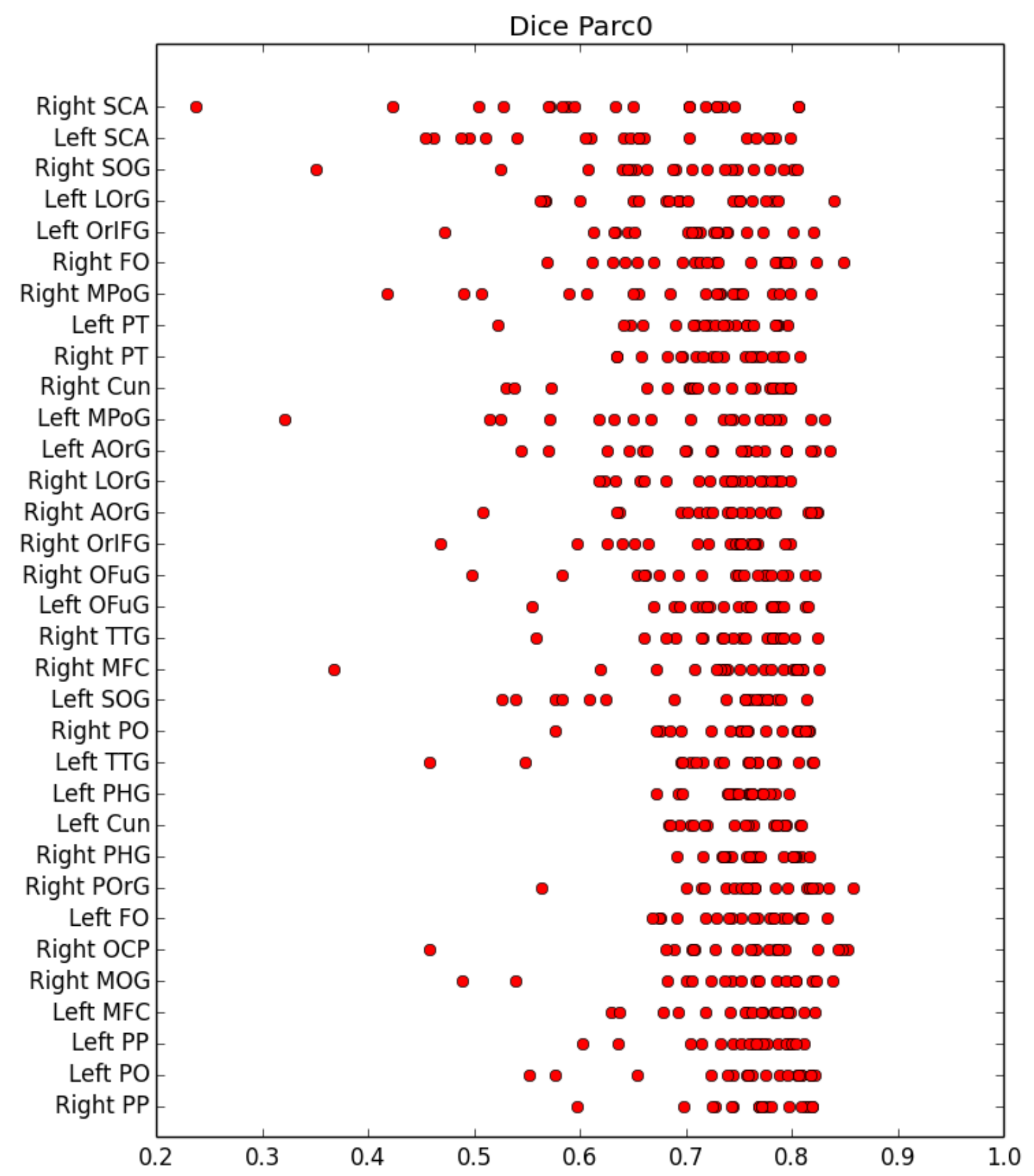
References

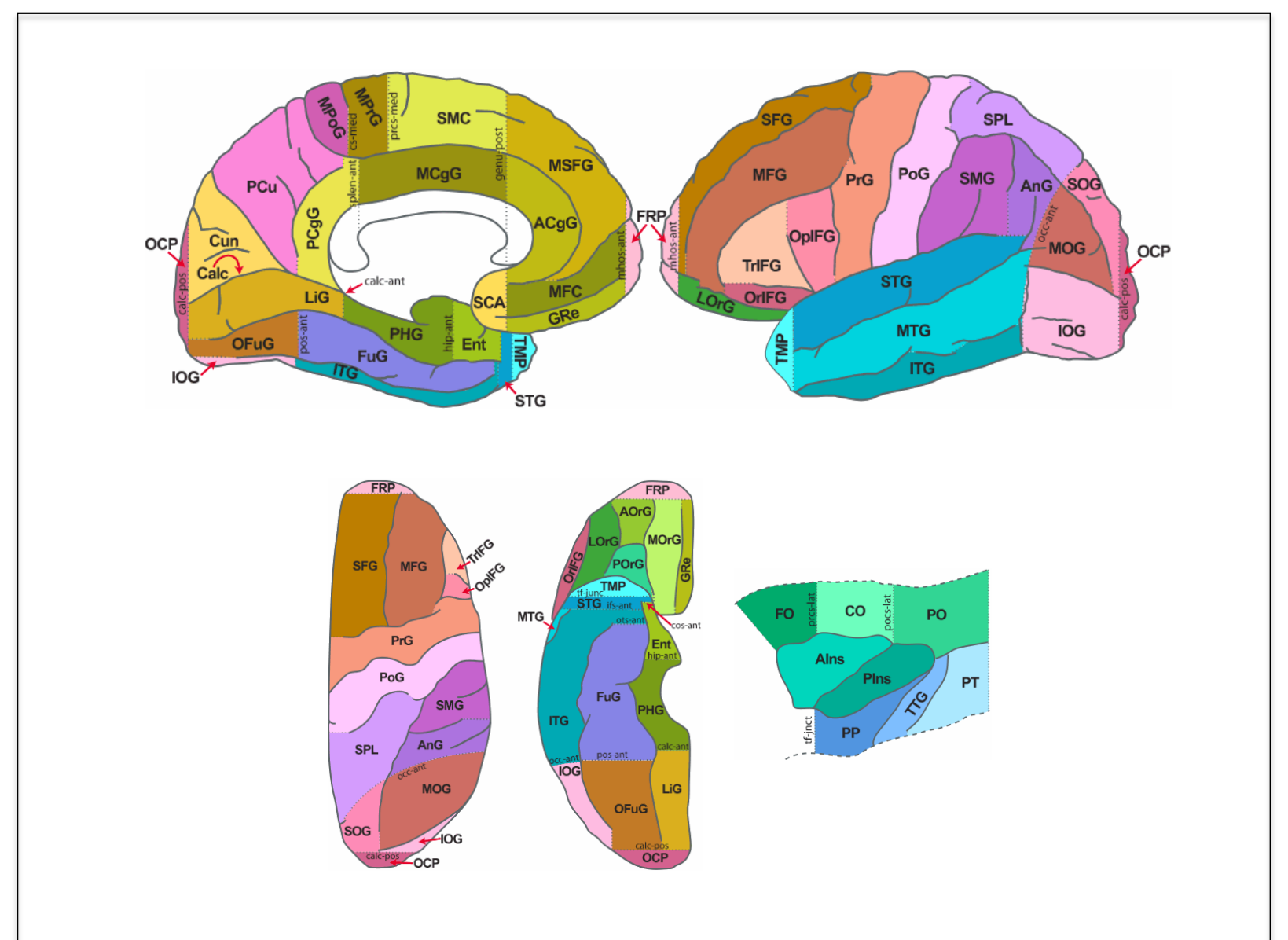
[1] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging. 2015;15:29. doi:10.1186/s12880-015-0068-x.

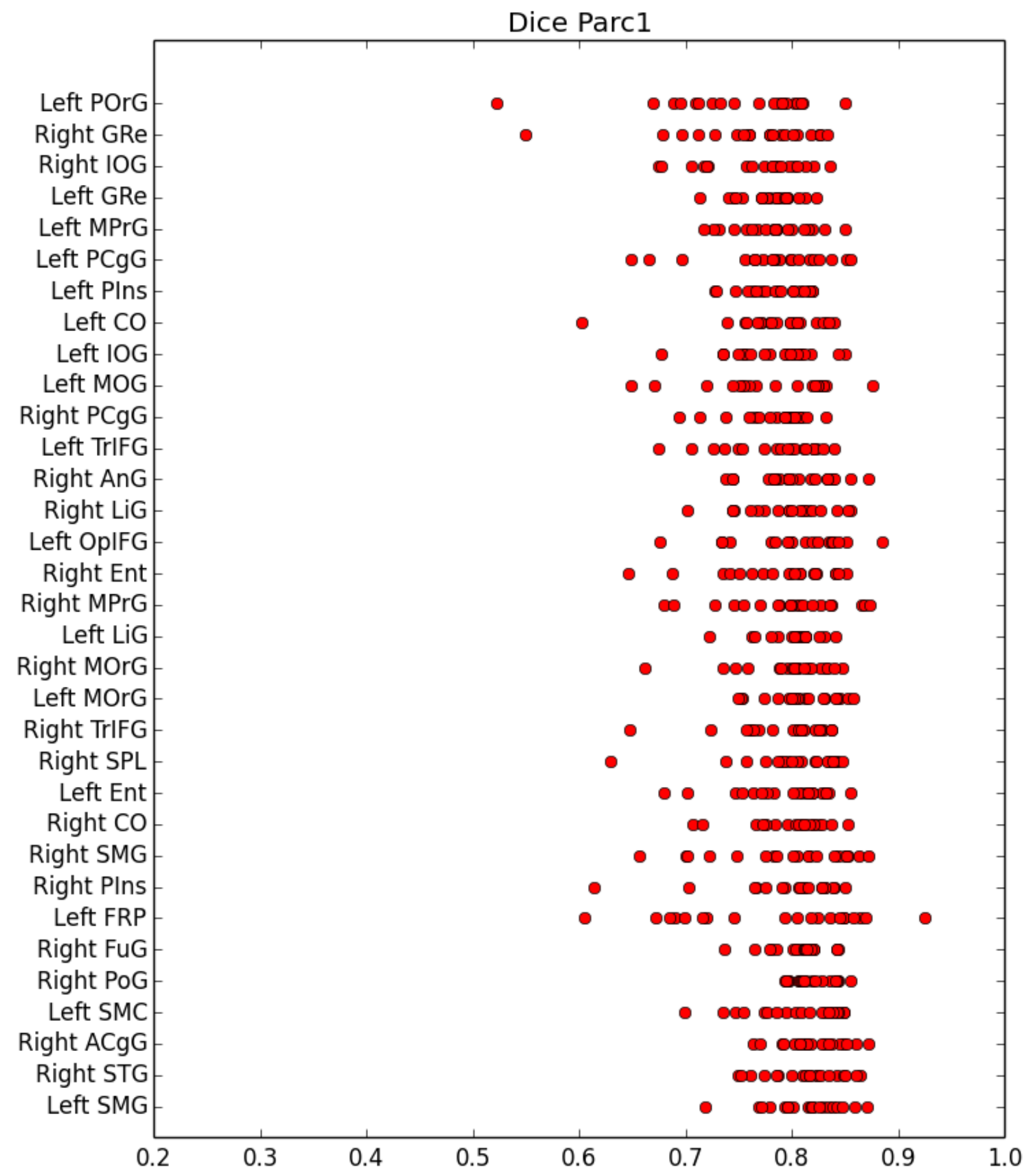
[2] Avants BB, Tustison NJ, Stauffer M, Song G, Wu B, Gee JC. The Insight ToolKit image registration framework. Frontiers in Neuroinformatics. 2014;8:44. doi:10.3389/fninf.2014.00044.

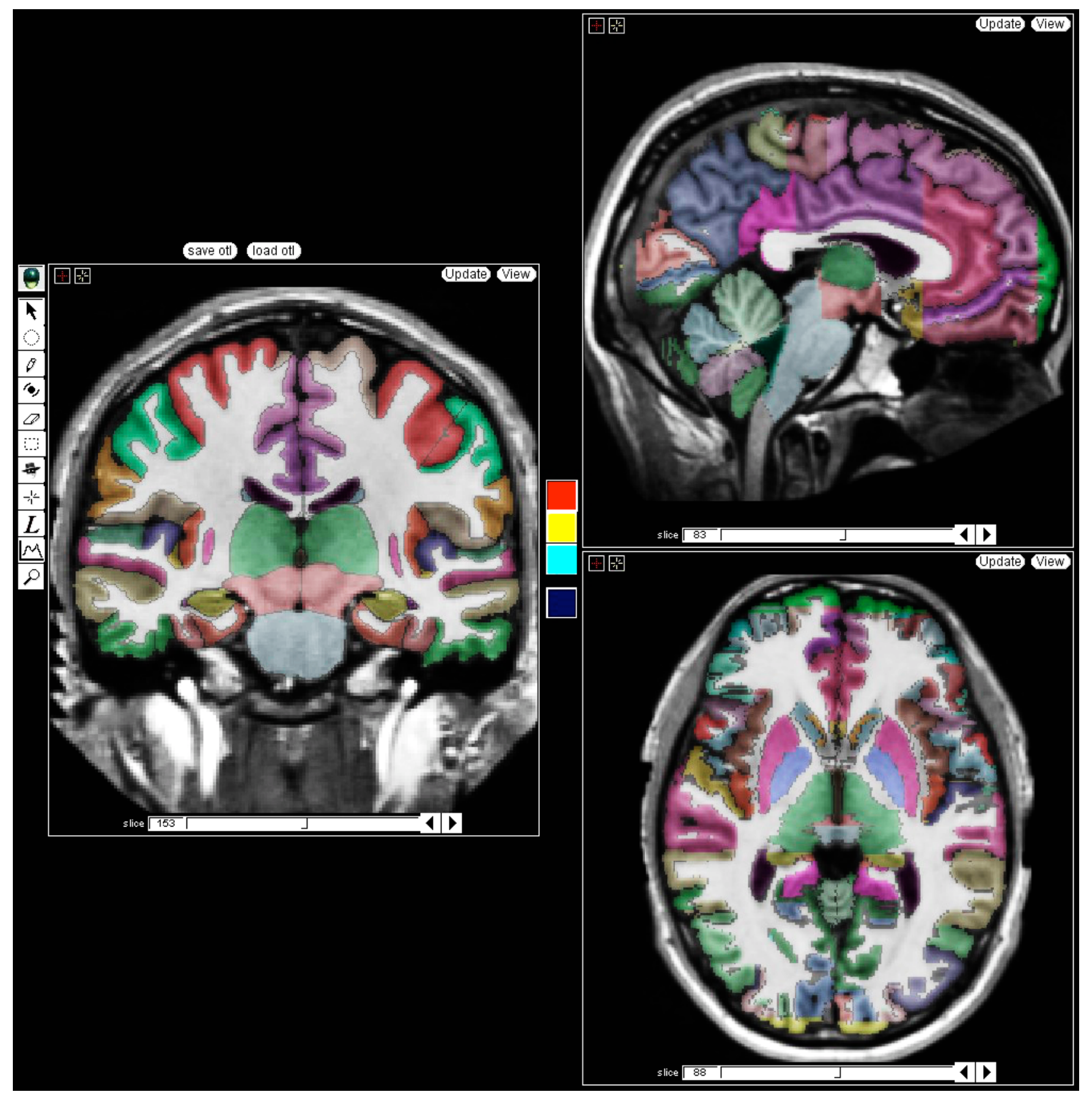
[3] Klein A, Andersson J, Ardekani BA, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage. 2009;46(3):786-802. doi:10.1016/j.neuroimage.2008.12.037.

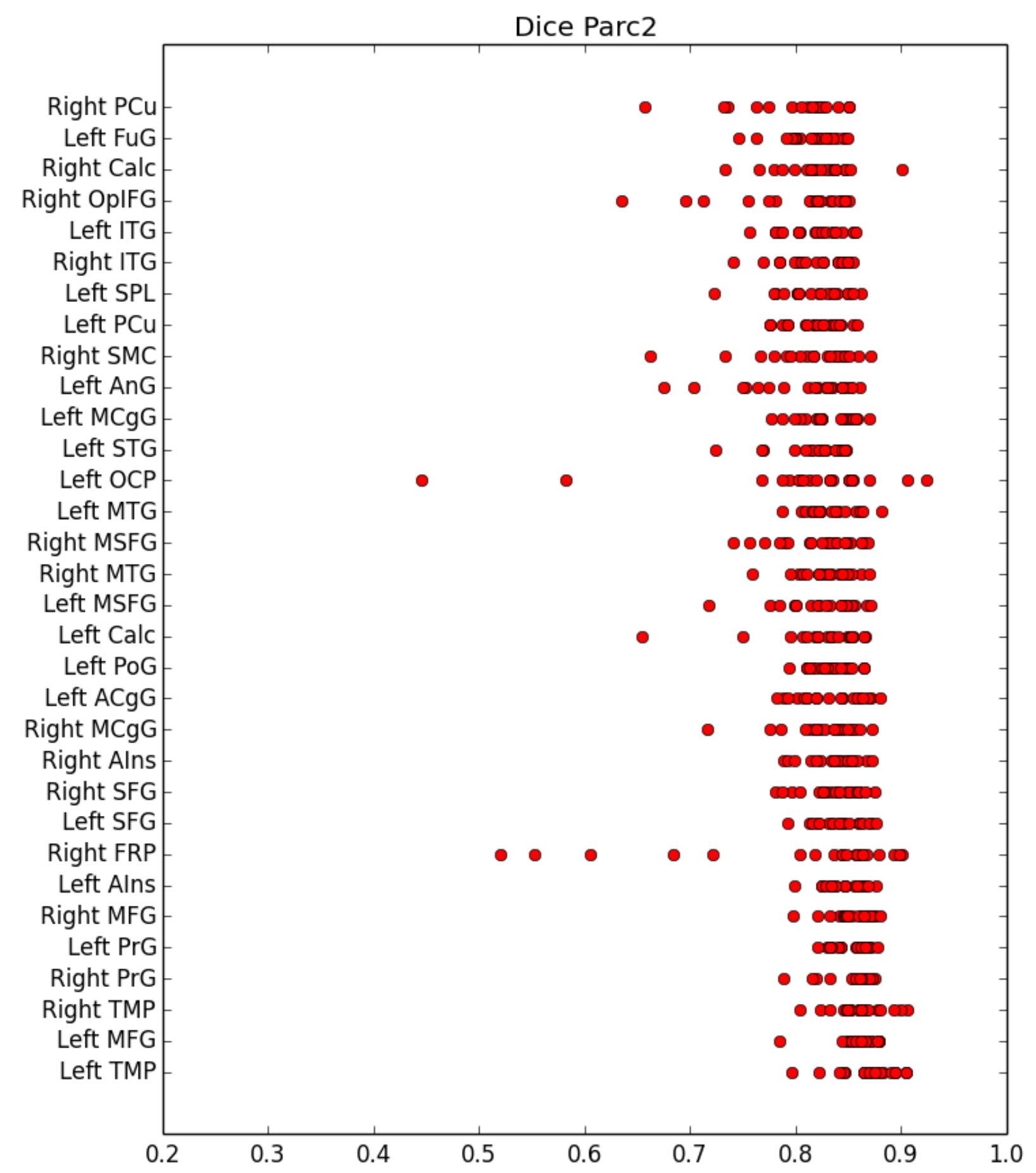





BrainCOLOR Cortical Labeling Protocol
http://braincolor.mindboggle.info/protocols/


Dice Parc1
Left POrG
Right GRe
Right IOG
Left GRe
Left MPrG
Left PCgG
Left PIns
Left CO
Left IOG
Left MOG
Right PCgG
Left TrIFG
Right AnG
Right LiG
Left OpIFG
Right Ent
Right MPrG
Left LiG
Right MOrG
Left MOrG
Right TrIFG
Right SPL
Left Ent
Right CO
Right SMG
Right PIns
Left FRP
Right FuG
Right PoG
Left SMC
Right ACgG
Right STG
Left SMG


Update
View


Dice Parc2
Right PCu
Left FuG
Right Calc
Right OpIFG
Left ITG
Right ITG
Left SPL
Left PCu
Right SMC
Left AnG
Left MCgG
Left STG
Left OCP
Left MTG
Right MSFG
Right MTG
Left MSFG
Left Calc
Left PoG
Left ACgG
Right MCgG
Right AIns
Right SFG
Left SFG
Right FRP
Left AIns
Right MFG
Left PrG
Right PrG
Right TMP
Left MFG
Left TMP