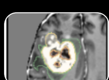
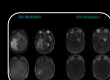
AI for Predicting Response

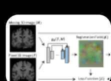
Jayashree Kalpathy-Cramer, PhD
(Ken Chang, Andrew Beers, Jay Patel, Katharina Hoebel, Elizabeth Gerstner, Bruce Rosen)
Athinoula A. Martinos Center for Biomedical Imaging
MGH/Harvard Medical School
AAPM 2019

Machine learning and AI in oncology


Segmentation


Response Assessment


Radiogenomics


Registration


Survival Prediction

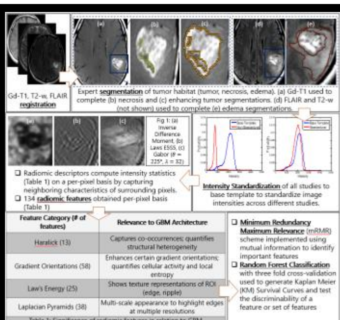

Drug delivery

Machine Learning

- Radiomics and deep learning are currently popular
- Radiomics typically involves developing classifiers using hand-crafted features of segmented regions
- Deep learning approaches typically use the whole image (or a bounding box around the ROI, without segmentation) to learn the filters

Typical radiomics workflow

Image acquisition → [Preprocessing] → [Tumor segmentation] → Feature extraction → Classification / stratification


Gd-T1, T2-w, FLAIR **registration**

Expert **segmentation** of tumor habitat (tumor, necrosis, edema). (a) Gd-T1 used to complete (b) necrosis and (c) enhancing tumor segmentations. (d) FLAIR and T2-w (not shown) used to complete (e) edema segmentations.

Fig 1: (a) Inverse Difference Moment, (b) Laws E5S5, (c) Gabor (θ = 225°, λ = 32)

- Radiomic descriptors compute intensity statistics (Table 1) on a per-pixel basis by capturing neighboring characteristics of surrounding pixels.
- 134 **radiomic features** obtained per-pixel basis (Table 1)

Intensity Standardization of all studies to base template to standardize image intensities across different studies.

Feature Category (# of features)	Relevance to GBM Architecture
Haralick (13)	Captures co-occurrences; quantifies structural heterogeneity
Gradient Orientations (58)	Enhances certain gradient orientations; quantifies cellular activity and local entropy
Law's Energy (25)	Shows texture representations of ROI (edge, ripple)
Laplacian Pyramids (38)	Multi-scale appearance to highlight edges at multiple resolutions

Table 1: Significance of radiomic features in relation to GBM

- **Minimum Redundancy Maximum Relevance** (mRMR) scheme implemented using mutual information to identify important features
- **Random Forest Classification** with three fold cross-validation used to generate Kaplan Meier (KM) Survival Curves and test the discriminability of a feature or set of features

Slide: Jay Patel

Shape / Morphology Features

Shape features are derived from areas, volume, and diameters of segmentation (mask)

These are irrespective of intensity and adjusted for voxel dimensions.

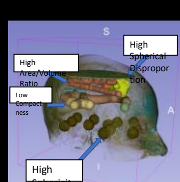
$$surface\ to\ volume\ ratio = \frac{A}{V}$$

$$spherical\ disproportion = \frac{A}{4\pi R^2} = \frac{A}{\sqrt[3]{36\pi V^2}}$$

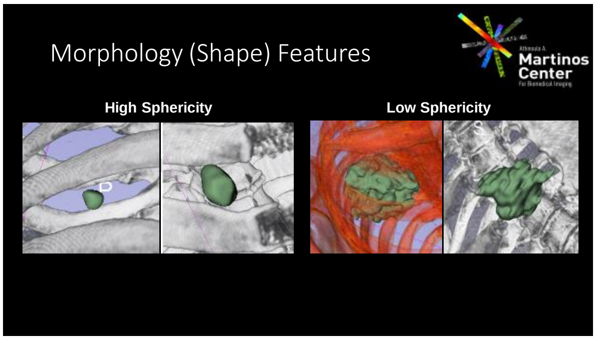
$$compactness\ 1 = \frac{V}{\sqrt{\pi A^3}}$$

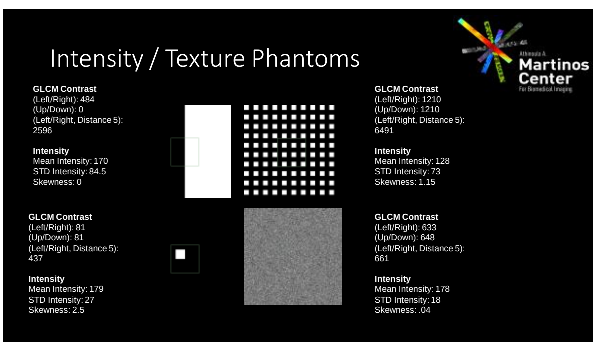
$$sphericity = \frac{\sqrt[3]{36\pi V^2}}{A}$$

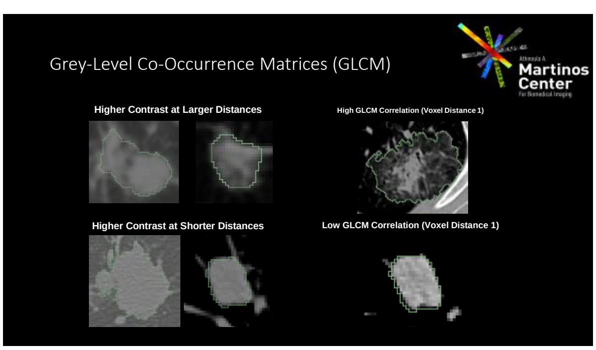
https://github.com/QTIM-Lab/qtim_tools/blob/master/documentation/Tutorial_Modules/Machine_Learning_and_Feature_Extraction_Tutorial/Feature_Extraction_Tutorial/Feature_Extraction_Tutorial.ipynb

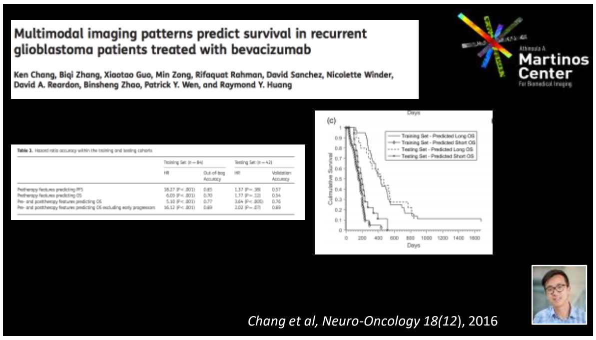


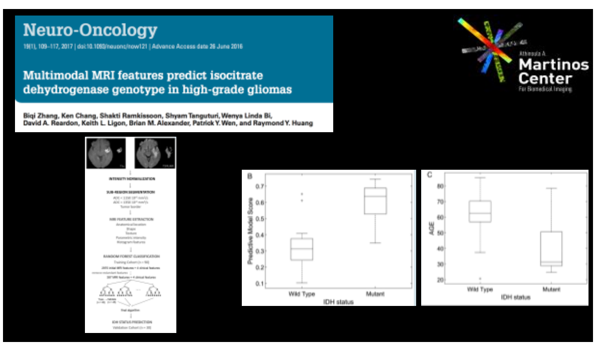
Slide: Andrew Beers

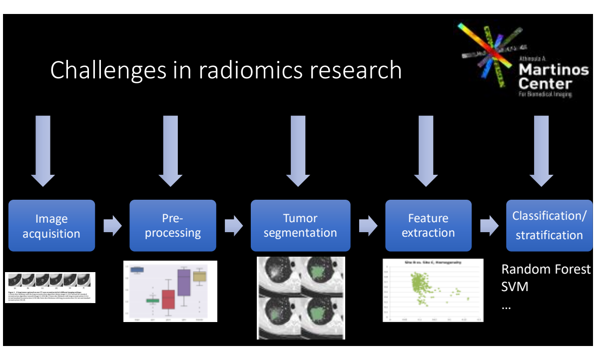

Morphology (Shape) Features
Martinos Center
High Sphericity
Low Sphericity


Intensity / Texture Phantoms
Martinos Center
GLCM Contrast
(Left/Right): 484
(Up/Down): 0
(Left/Right, Distance 5): 2596
Intensity
Mean Intensity: 170
STD Intensity: 84.5
Skewness: 0
GLCM Contrast
(Left/Right): 1210
(Up/Down): 1210
(Left/Right, Distance 5): 6491
Intensity
Mean Intensity: 128
STD Intensity: 73
Skewness: 1.15
GLCM Contrast
(Left/Right): 81
(Up/Down): 81
(Left/Right, Distance 5): 437
Intensity
Mean Intensity: 179
STD Intensity: 27
Skewness: 2.5
GLCM Contrast
(Left/Right): 633
(Up/Down): 648
(Left/Right, Distance 5): 661
Intensity
Mean Intensity: 178
STD Intensity: 18
Skewness: .04


Grey-Level Co-Occurrence Matrices (GLCM)
Martinos Center
Higher Contrast at Larger Distances
High GLCM Correlation (Voxel Distance 1)
Higher Contrast at Shorter Distances
Low GLCM Correlation (Voxel Distance 1)


Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab
Ken Chang, Biqi Zhang, Xiaotao Guo, Min Zong, Rifaquat Rahman, David Sanchez, Nicolette Winder, David A. Reardon, Binsheng Zhao, Patrick Y. Wen, and Raymond Y. Huang
Martinos Center
For Biomedical Imaging
Training Set - Predicted Long OS
Training Set - Predicted Short OS
Testing Set - Predicted Long OS
Testing Set - Predicted Short OS
Cumulative Survival
Days
Chang et al, Neuro-Oncology 18(12), 2016


Neuro-Oncology
19(1), 109–117, 2017 | doi:10.1093/neuonc/now121 | Advance Access date 26 June 2016
Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas
Biqi Zhang, Ken Chang, Shakti Ramkissoon, Shyam Tanguturi, Wenya Linda Bi, David A. Reardon, Keith L. Ligon, Brian M. Alexander, Patrick Y. Wen, and Raymond Y. Huang
Martinos Center
For Biomedical Imaging
Predictive Model Score
AGE
Wild Type
Mutant
IDH status


Challenges in radiomics research
Martinos Center
For Biomedical Imaging
Image acquisition
Pre-processing
Tumor segmentation
Feature extraction
Classification/ stratification
Random Forest
SVM
…

Acquisition effects

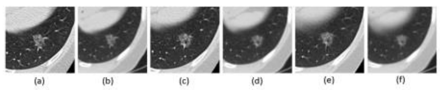


Figure 1. A lung tumor captured on one CT scan reconstructed at 6 different imaging settings: 1.25mm slice thickness with the lung reconstruction algorithm (sharp image) (1.25L) (**a**) and the standard reconstruction algorithm (smooth image) (1.25S) (**b**); 2.5mm slice thickness with lung reconstruction (2.5L) (**c**) and standard reconstruction (2.5S) (**d**); 5mm slice thickness with lung reconstruction (5L) (**e**) and standard reconstruction (5S) (**f**).

Zhao et al https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4806325/pdf/srep23428.pdf

Effects of image reconstruction

CCC heat map of radiomic features. The CCCs (0 to 1) of the studied radiomic features were computed from repeat CT images reconstructed at (a) six identical imaging settings or (b) three different imaging settings. There were 89 quantitative features grouped into 15 feature classes. The brighter the red color, the higher the CCC value (i.e., the more reproducibility) of a feature computed for the repeat scans. The label of "1.25L1 vs 1.25L2" means both first and second scans were reconstructed at 1.25mm slice thickness using the lung algorithm. "2.5L vs 2.5S" means both scans were reconstructed at 2.5mm slice thickness but using different algorithms (i.e., lung vs. standard algorithms).

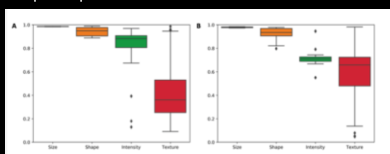
Zhao et al https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4806325/pdf/srep23428.pdf

Radiomics in MR – a repeatability study

- Radiomics used commonly in MR
  - However, MR is often not quantitative ("T1-weighted")
  - Normalization matters enormously
  - Preprocessing matters
  - Implementation matters
- ICC of many features <0.5

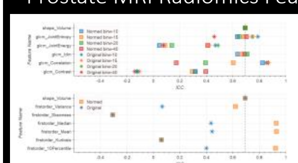
Additional challenges in MRI

- Pixel intensity in “weighted” images does not have inherent meaning
- Often multiple sequences are used

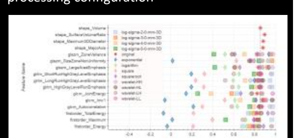


Repeatability of radiomics features using open source software for T1 and FLAIR ROI in GBM

Repeatability of Multiparametric Prostate MRI Radiomics Features


(a) Tumor ROI (top: texture features, bottom: first order features)

“Our study shows that radiomics features...vary greatly in their repeatability. Furthermore, repeatability of radiomics features evaluated using ICC is highly susceptible to the processing configuration”


(a) T2w Tumor ROI

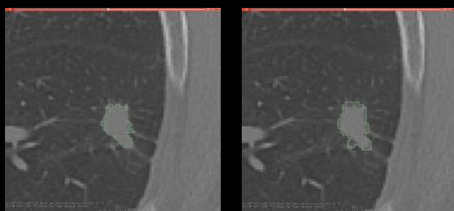
Schweir et al, https://arxiv.org/pdf/1807.06089.pdf

Repeatability and Reproducibility of Radiomic Features: A Systematic Review

“Investigations of feature repeatability and reproducibility are currently limited to a small number of cancer types. No consensus was found regarding the most repeatable and reproducible features with respect to different settings.”

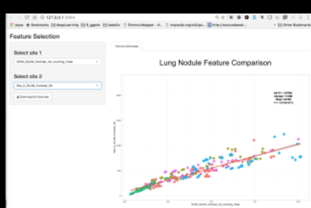
Traverso et al, 2018, International Journal of Radiation Oncology* biology* physics

Effect of segmentation on texture



High contrast 5 vs 10

Lack of reproducibility: Implementations can agree between packages



- Contrast

Or not…

- Contrast

Standardization in Quantitative Imaging: A Comparison of Radiomics Feature Values Obtained by Different Software Packages On a Set of Digital Reference Objects

By computing a subset of nine common radiomics features using a variety of software packages on DROs, we have shown that while several features agree strongly, others do not. This highlights the need for standardization in feature definitions and proof of equivalence of computational methods.

Tuesday, 7/16/2019) 4:30 PM - 6:00 PM, Room: 304ABC AAPM 2019

Radiotherapy and Oncology 130 (2019) 2–9

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

ELSEVIER

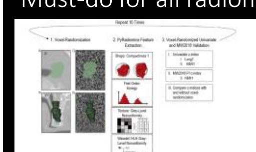
Original article

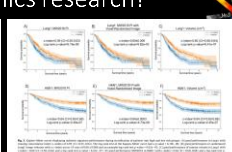
Vulnerabilities of radiomic signature development: The need for safeguards

Mattea L. Welch [a,f,i], Chris McIntosh [e,f,i], Benjamin Haibe-Kains [a,c,i,j], Michael F. Milosevic [b,e,i], Leonard Wee [g], Andre Dekker [g], Shao Hui Huang [b,i], Thomas G. Purdie [b,e,f,i], Brian O'Sullivan [b,i], Hugo J.W.L. Aerts [h], David A. Jaffray [a,b,d,e,f,i,*]

[a] Department of Medical Biophysics, University of Toronto; [b] Department of Radiation Oncology, University of Toronto; [c] Ontario Institute of Cancer Research, Toronto; [d] IBBME, University of Toronto; [e] Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto; [f] The Techna Institute for the Advancement of Technology for Health, Toronto, Canada; [g] Department of Radiation Oncology (MAASTRO), GROW Research Institute, Maastricht University, the Netherlands; [h] Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, USA; [i] Princess Margaret Cancer Centre, University Health Network; and [j] Vector Institute, Toronto, Canada

Must-do for all radiomics research!

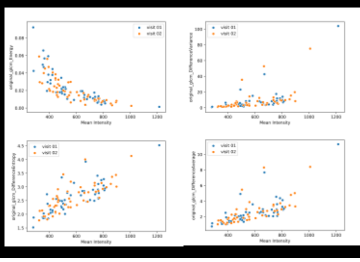




"MW2018 had an external validation concordance index of 0.64. However, a similar performance was achieved using features extracted from images with randomized signal intensities (c-index=0.64 and 0.60 for H&N and lung respectively). Tumour volumes had a c-index =0.64 and correlated strongly with three of the four model features. It was determined that the signature was a surrogate for tumour volume and that intensity and texture values were not pertinent for prognostication"

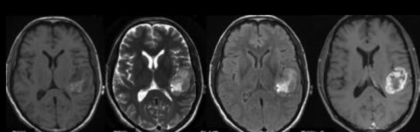
Welch et al, 2019

Features are highly correlated (to volume and intensity!)



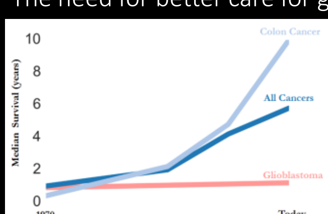
Glioblastoma

- GBM is an aggressive form of brain cancer
- It affects people of all ages, and it carries a poor prognosis
- Standard of care is chemotherapy, radiotherapy, and surgery



Source: Mahajan et al., Clinical Radiology 2015

The need for better care for glioma



Macmillan Cancer Support (2011)
Alexander et al., JCO (2017)

5-year survival rate after diagnosis: 5%

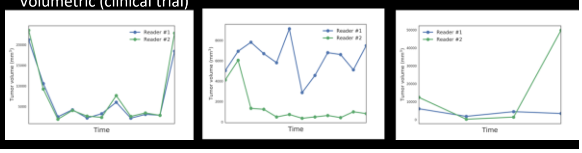
Variability in response assessment

RANO

Moderate agreement

(adjudication rates in clinical trials is often high)

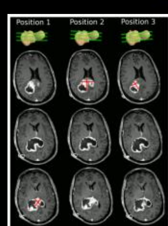
Volumetric (clinical trial)



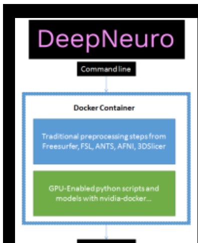
Effect of head tilt on (human) measurements of tumor burden

Reuter et al analyzed the reliability of the area measure with respect to head placement in the MRI scanner and compares it with 3D volume measures in a dataset of 8 subjects

	2D RANO	3D Vol Thick	3D Vol Thin
RMSE	22%	4.8%	3.2%
ICC	0.93	0.991	0.998
MLPI	59%, 25%	12%	8%



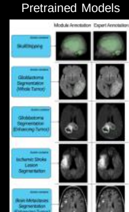
Reuter et al, J Neurooncol (2014) 118:123–129



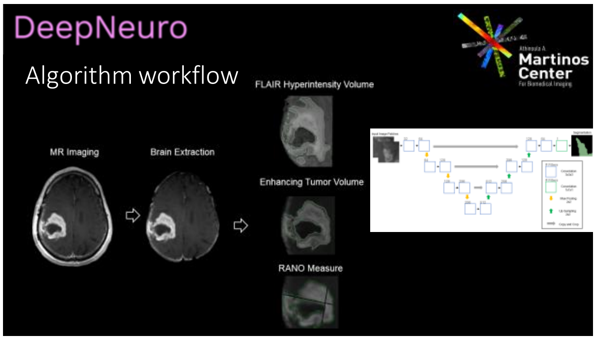
Pipeline Customization

Image Preprocessing

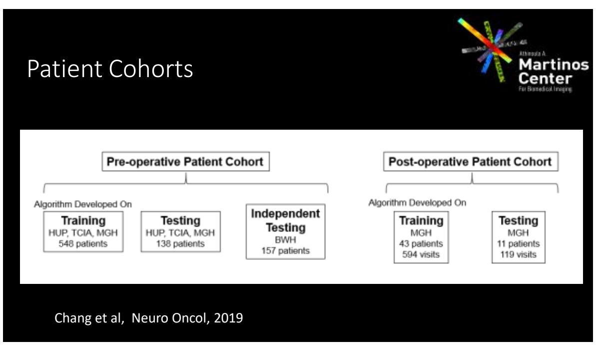
Pretrained Models



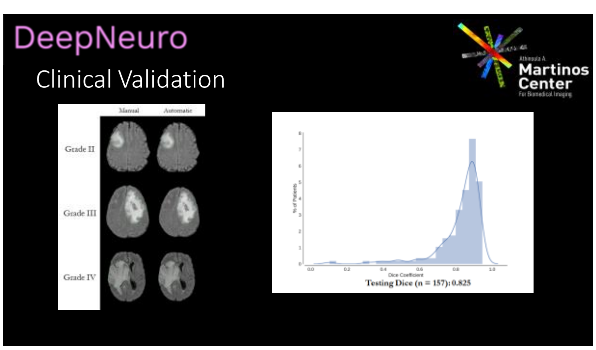
https://github.com/QTIM-Lab/DeepNeuro



DeepNeuro
Algorithm workflow
FLAIR Hyperintensity Volume
MR Imaging
Brain Extraction
Enhancing Tumor Volume
RANO Measure
Martinos Center

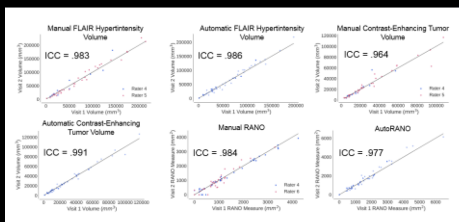


Patient Cohorts
Martinos Center
Pre-operative Patient Cohort
Algorithm Developed On
Training
HUP, TCIA, MGH
548 patients
Testing
HUP, TCIA, MGH
138 patients
Independent Testing
BWH
157 patients
Post-operative Patient Cohort
Algorithm Developed On
Training
MGH
43 patients
594 visits
Testing
MGH
11 patients
119 visits
Chang et al, Neuro Oncol, 2019

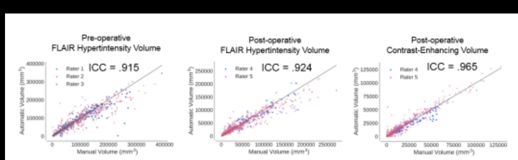


DeepNeuro
Clinical Validation
Martinos Center
Manual
Automatic
Grade II
Grade III
Grade IV
% of Patients
Dice Coefficient
Testing Dice (n = 157): 0.825

Both manual and automatic measures are highly repeatable



High agreement between manual and automatic volume



Glioblastoma response assessment

- Longitudinal measurement of glioma burden with MRI is the basis for treatment response assessment
- Manual delineation of tumor burden is both time-consuming and subject to inter-rater variability
- We developed a deep learning algorithm that automatically segments abnormal FLAIR hyperintensity and contrast-enhancing tumor, quantitating tumor volumes as well as the product of maximum bi-dimensional diameters according to the Response Assessment in Neuro-Oncology (RANO) criteria (AutoRANO)

Response Assessment in Brain Tumors

- Goal: Used to assess response in brain tumors
  - Complete Response
  - Partial Response
  - Stable Disease
  - Progressive Disease
- Adjudication rates are quite high in oncology clinical trials (~40%)

	CR	PR	SD	PD
T1-Gd+	None	≥50%	<50%↓ - <25%↑	≥25%↑*
T2/FLAIR	Stable or ↓	Stable or ↓	Stable or ↓	↑*
New Lesion	None	None	None	Present*
Corticosteroids	None	Stable or ↓	Stable or ↓	NA
Clinical Status	Stable or ↑	Stable or ↑	Stable or ↑	↓*
Requirement for Response	All	All	All	Any*

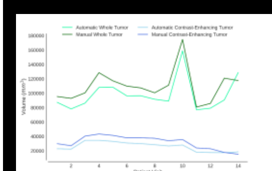
CR = Complete Response, PR = Partial Response, SD = Stable Disease, PD = Progressive Disease

https://www.omicsonline.org/open-access/adjudication-rates-between-readers-in-blinded-independent-central-review-ofoncology-studies-2167-0870-1000289.php?aid=81626

Treatment Response Assessment

Aim 1 Aim 2 Aim 3

Longitudinal Tracking of Tumor Burden



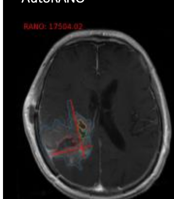
Can our **automatically derived RANO measurements** be in agreement with human measurements?

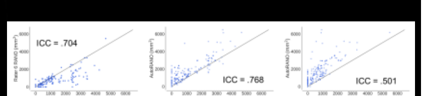
And be in agreement with treatment response assessment in the following categories?

- Complete Response
- Partial Response
- Stable Disease
- Progressive Disease

DeepNeuro

AutoRANO

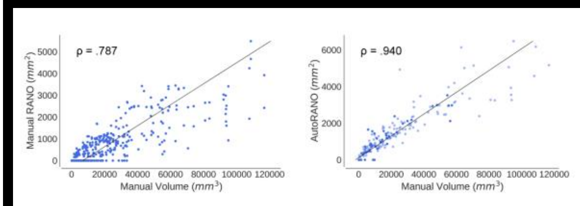




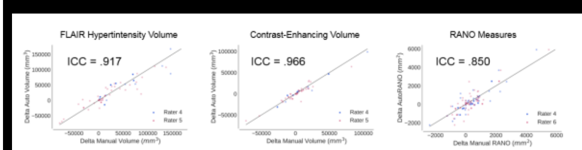
Comparison between experts, AutoRANO to expert A, AutoRANO to expert B

Chang et al, Neuro-Oncology (2019)

AutoRANO is more correlated with manual volume than Manual RANO

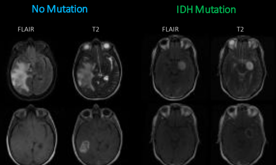


Both automatic volume and AutoRANO can track longitudinal changes in tumor burden

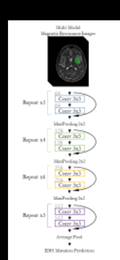


DeepNeuro

IDH prediction



Less aggressive growth
Sharp margins
Homogenous signal intensity
Less contrast enhancement

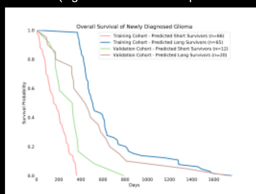


	Accuracy
Training	93%
Validation	94%
Testing	88%

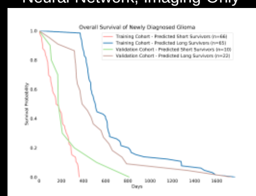
Chang et al, Clinical Cancer Research

Prediction of Overall Survival

Radiomics (Age + Volumes + Shape + Texture)



Neural Network, Imaging Only



DeepNeuro

Brain metastases

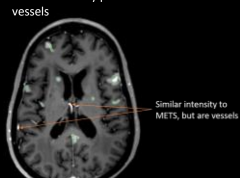
- Brain metastases (BM) patients undergo routine MR scans throughout therapy
- Need to track individual lesion growth/shrinkage rates across timepoints to assess efficacy of current treatment regimen
- Manual delineation of entire lesion burden is too time-consuming to be feasible in clinical workflow

- Solution: utilize neural network to segment lesions on MPRAGE-post contrast imaging

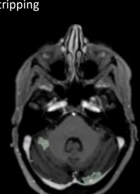
DeepNeuro METS Segmentation

Challenges

- Large number of micrometastases (>10 per patient on average)
- Similar intensity profile of METS and vessels

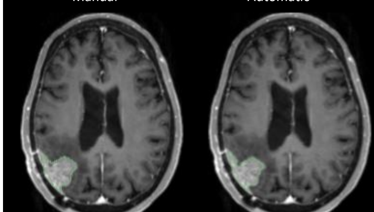


- Dural lesions can be accidentally removed from automatic skull stripping



Results – Loss Functions

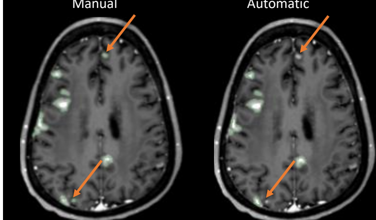
Manual Automatic



	Training	Validation	Testing
Dice Loss	0.72	0.65	0.65
Weighted Cross-Entropy	0.76	0.69	0.69
Boundary weighted Cross-Entropy	**0.78**	**0.74**	**0.70**

Results – Detection Rates

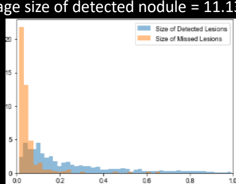
Manual Automatic

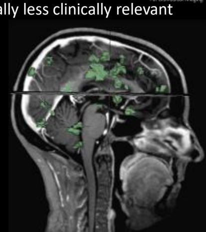


	Average Size of Detected Lesions (mL)	Average Size of Missed Lesions (mL)
Training	11.13	0.10
Validation	6.52	0.07
Testing	4.06	0.05

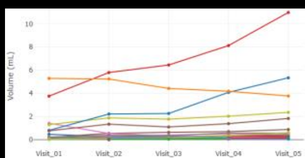
Results – Detection Rates

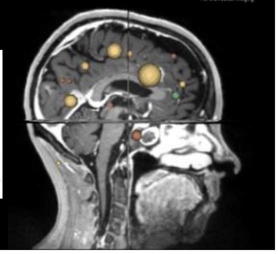
- Micrometastases hard to detect, but potentially less clinically relevant
- Average size of missed nodule = .09 mL
- Average size of detected nodule = 11.13 mL



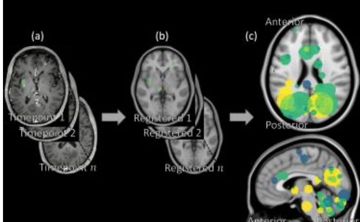


Longitudinal Tracking of Brain METS



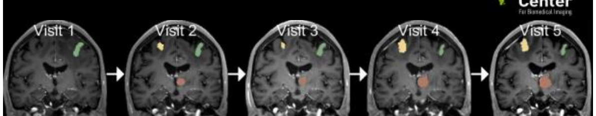


Longitudinal Tracking of Brain METS



(a) Deep learning based approach used to automatically segment metastatic lesions across time points. (b) Affine registration of all patients/all timepoints to MNI atlas. Lesion segmentations are subsequently projected onto the atlas space. (c) Lesion growth rate over timepoints is measured and overlaid on atlas. Circle size corresponds to average lesion size over timepoints and color corresponds to growth rate (blue: shrinkage; green: little to no growth; yellow: high growth). Significantly larger and higher growth rate metastases found in posterior anatomy.

Longitudinal Tracking of Brain METS



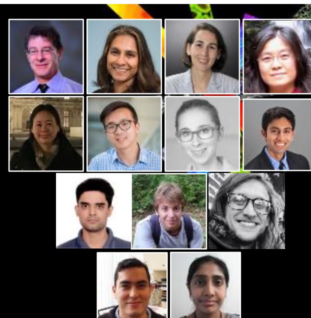
Red lesion represents high growth rate metastases. This lesion was not present during time point one, but grows to a significant size over the patient visits. Yellow lesion represents slow growth rate metastases. This lesion was also not present during time point one, but does not grow to as large of a volume, nor grow as quickly as the red-labeled lesion. Green lesion represents a shrinking metastases. This lesion shrinks in size over the course of treatment.

Summary

- Machine learning including deep learning has great potential in response assessment.
- However, the repeatability and reproducibility of radiomics is an area of active research.
- Deep learning based approaches for segmentation and registration have demonstrated good performance in the literature
- Deep learning methods can be brittle and not generalize well
- Deep learning methods are considered to be "black boxes" but techniques are being developed for explainable AI
- Comparing sophisticated models to baseline volume change is highly recommended

Acknowledgements

- Athinoula A. Martinos Center for Biomedical Imaging
  - Jayashree Kalpathy-Cramer
  - Elizabeth Gerstner
  - Bruce Rosen
  - Yi-Fen Yen
  - Ina Ly
  - Andrew Beers
  - Ken Chang
  - Katharina Hobel
  - Jay Patel
  - Jonathan Cardona
  - Praveer Singh
  - Malika Shahrawat
  - Sunakshi Paul
- Funding & Support
  - MGH/BWH Center for Clinical Data Science
  - National Science Foundation
  - National Institutes of Health



qtim-lab.github.io