Robustness Of Radiomics To Variations In Segmentation Methods In Multimodal Brain MRI

Maarten Gijsbert Poirot (m.g.poirot@amsterdamumc.nl)
Amsterdam University Medical Center
Matthan Caan
Amsterdam University Medical Center
Henricus Gerardus Ruhe
Amsterdam University Medical Center
Atle Bjørnerud
University of Oslo
Inge Groote
Oslo University Hospital
Liesbeth Reneman
Amsterdam University Medical Center
Henk Marquering
Amsterdam University Medical Center

Article

Keywords:

Posted Date: March 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1444224/v1

Abstract

Background: Radiomics in neuroimaging uses fully automatic segmentation to delineate the anatomical areas for which radiomic features are computed. However, differences among these segmentation methods affect radiomic features to an unknown extent.

Method: A scan-rescan dataset (n=46) of T1-weighted and diffusion tensor images was used. Subjects were split into a sleep-deprivation and a control group. Scans were segmented using four segmentation methods from which radiomic features were computed. First, we measured segmentation agreement using the Dice-coefficient. Second, robustness and reproducibility of radiomic features were measured using the intraclass correlation coefficient (ICC). Last, difference in predictive power was assessed using the Friedman-test on performance in a radiomics-based sleep deprivation classification application.

Results: Segmentation agreement was generally high (interquartile range=0.77-0.90) and median feature robustness to segmentation method variation was higher (ICC>0.7) than scan-rescan reproducibility (ICC 0.3-0.8). However, classification performance differed significantly among segmentation methods ($p<0.001$) ranging from 77% to 84%. Accuracy was higher for more recent deep learning-based segmentation methods.

Conclusion: Despite high agreement among segmentation methods, subtle differences significantly affected radiomic features and their predictive power. Consequently, the effect of differences in segmentation methods should be taken into account when designing and evaluating radiomics-based research methods.

1. Introduction

Radiomics is an established method for quantitative analysis of radiological images. It involves processing of medical images to extract large numbers of quantitative image features.[1–3] In clinical oncology, radiomics has (greatly) contributed to prediction of patient outcome and clinical decision-making support.[4] This success has increased interest for applications of radiomics in other disciplines, such as psychiatry. The application of radiomics for psychiatric disorders constitutes to a relatively new field of research: psychoradiology[5–7].

Radiomics in psychiatry has seen several applications such as classification, prediction, and treatment selection[5,7] in diseases like schizophrenia[8–12], attention hyperactivity disorder[13], bipolar disorder[14], and major depressive disorder[15,16]. In these applications, various magnetic resonance imaging (MRI) modalities have been employed such as structural T1-weighted imaging, T2-FLAIR-weighted imaging, diffusion tensor imaging (DTI), functional MRI (fMRI), and arterial spin labeling (ASL).

Despite these promising applications, reliability of radiomics has been hindering its broad validity and generalizability. Variability and uncertainty can be introduced in each step in the six-step radiomics pipeline: First, image acquisition[17–25], processing[24,26,35,27–34] and segmentation[2,3,36–40], then feature

extraction and selection, and finally statistical inference[1,3,36,37,41,42]. Reliability of radiomics with respect to image acquisition and image processing has been under extensive scientific scrutiny, especially in oncology, where 87% of studies on robustness of radiomic features has been performed.[43] This research has yielded standardization of MRI-acquisition protocols and standardization of radiomics definitions that the field of psychoradiology can draw from.[44,45] However, the segmentation in oncology has mainly been focused on effects of semi-automatic segmentation[34,46–51] of tumor mass rating, whereas psychoradiology makes use of different fully-automatic whole-brain anatomical segmentation methods.[52] Thus, oncological findings on robustness to segmentation methods are hard to translate to psychoradiology.

Second, domain specific research focusing on the robustness of brain MRI radiomic features for psychoradiology has been lacking. To our knowledge, only one study has been published in this respect. In this study, Li et al.[53] found that texture features are the most reproducible brain MRI features in T1-weighted images for hippocampal segmentations. Whole-brain and subcortical segmentation accuracy was assessed in comparative settings by several authors[54–56] Other studies that investigated segmentation accuracy were limited to one or two anatomical regions, such as the hippocampus,[57–62] amygdala[60] and caudate and putamen[63], but have not reported effects on radiomic features.

In this work, we analyze the impact of variations in segmentation methods on a full radiomics pipeline that uses DTI and T1-weighted radiomic features for classification purposes. Segmentation methods included are two established methods: FreeSurfer SAMSEG[65,66] and FreeSurfer ASEG[64]. Two other methods included are recent deep learning-based methods: FastSurfer[55] and. We analyze this full radiomics pipeline in three subsequent steps:

1) We analyzed segmentation agreement between pairs of segmentation methods to aid interpretation of following robustness findings.

2) We analyzed robustness of radiomic features to segmentation method variation by computing the ICC across segmentation methods for each radiomic feature. We compare these numbers to scan-rescan reproducibility of radiomic features.

3) We compared discriminative power of the radiomic features generated by these four segmentation methods by subjecting them to a diagnostic test, consisting of classifying sleep-deprived subjects and non-deprived controls. Literature has previously shown subtle differences in changes in structural and diffusion weighted imaging after sleep deprivation.[67–71] For sake of dimensionality reduction and generalizability, classification was performed on a subset of all features, as is recommended in literature.[25,72–74]

2. Results

Of all 46 participants, data were present and consistent. No subjects had to be excluded for initial analysis. None of the segmentations nor DTI co-registrations failed. Thirty-two anatomical regions were available for all segmentation methods. Two examples of such segmentations are provided in Supplementary Fig. S1. Subsequently, radiomic feats were extracted from T1-weighted and DTI data over four time points. The total number of features extracted was 107, each of which can be attributed to one of the seven feature classes mentioned earlier.

2.1 Segmentation Agreement

Dice-coefficients were computed for each segmentation method pair. Fig. 1 shows segmentation agreement for each anatomical area, but left-right averaged where possible. Segmentation agreement as computed for each anatomical area was high (IQR = 0.77-0.90). Average agreement was highly correlated (ρ=0.93) between left and right anatomical areas, with the exception of the pallidum where left scored worse than right with a Dice-score difference of 0.1.

2.2 Radiomic Feature Reliability

Supplementary Fig. S2 provides an exhaustive overview of radiomic feature reliability as calculated for each feature, for each anatomical region, for each modality and according to the two earlier defined metrics: scan-rescan reproducibility and robustness to segmentation method variation. Remaining values were averaged over anatomical regions and feature class (Fig.2). The choroid plexus, inferior lateral ventricles and cerebrospinal fluid (CSF) were excluded due to subpar segmentation quality as prescribed in methods section 2.4.

2.3 Feature selection and classification

For the sleep deprivation classification, four subjects were excluded because of missing data, one of which was in the SWC group. Thus, the remaining cohort consisted of 22 subjects in the normal SWC group and 20 in the sleep-deprived group. The training partition consisted of 30 subjects, the validation partition of eight subjects, and the test set of four subjects.

Initially, about 27 thousand features were generated for each subject: 107 radiomic features, 32 anatomical regions, four time points and two modalities. A subset of features was selected for subsequent analysis in three steps described in section 2.5. In the first step, only first-order shape and first-order-based features of the DTI images were included. In the second, excluded anatomical regions are the choroid plexuses, the third and fourth ventricles and the CSF. After feature selection, only 640 features remained (32 radiomic features, 20 anatomical regions, one coefficient-over-time, one modality).

Computation time per fold amounted to be about a minute. Median performance results were computed both in terms of the training loss optimized (BCE) as well as classification accuracy (Table 1). The Friedman test yielded 31.18 (p value < 10^{-6}), rejecting the hypothesis that performance was similar among different pipelines.

3. Discussion

In this work, we have found that despite high whole-brain segmentation agreement, radiomic feature robustness to variability among these selected segmentation methods is only moderate. Consequently, this variability significantly affects radiomics-based classification performance.

As compared to literature, this work presents three novel analyses. First, this work presents a broader analysis of radiomic feature reliability than previously presented in literature, by including both a wide range of subcortical areas as well as two MRI modalities, i.e., T1- and diffusion weighted MRI.

Second, whereas existing literature is generally confined to one-to-one comparisons of methods, our work provides an independent comparison of four segmentation methods. The unbiased nature of this work counters a potential bias in favor of a newly introduced methods in study design. In addition, by including four segmentation methods in the same setup, this work provides a more comparable comparison among them.

Third, as opposed to most literature concerning comparison of sensitivity of whole-brain segmentation methods, our work takes a radiomic approach. We thereby go beyond other work testing sensitivity on volume measurements alone, such as in application to Alzheimer's disease.[55,56]

3.1. Segmentation agreement

In the first part of this work, we have analyzed the agreement between segmentation method pairs. Disagreement can stem from a variety of sources. For one, it can be anatomically dependent. Insufficient contrast of the specific anatomy in T1 can make certain anatomical regions hard to segment. In addition, a high surface-to-volume ratio can leave certain anatomical regions vulnerable to rapid reduction of Dice-coefficients. Non-anatomically inherent disagreement can be caused by differences in segmentation labeling definitions. This is the likely reason for the low Dice-coefficients in the inferior lateral ventricle and CSF produced by the SAMSEG method as compared to the other three methods. Additionally, differences in labeling definitions in atlases for conventional methods or training data for deep learning can be the root of systemic bias. This might be the cause for slightly reduced agreement between SAMSEG and the other methods, e.g., for cortical GM and WM. Thus, all methods showed high agreement, with only SAMSEG showing slight deviation.

3.2. Reliability of radiomic features

The second part of this work concentrated on the effects on radiomic feature reliability, broken down into scan-rescan reproducibility and robustness to variations in segmentation method. We discuss these reliability results along the lines of causes of modality, radiomic feature (class), and anatomical region.

Variations in radiomics measures using the different segmentation methods were lower than scan-rescan variability using any given segmentation method. Robustness in DTI consistently outperformed T1-weighted imaging, but this does not hold true for scan-rescan reproducibility. A potential cause of DTI

outperforming T1-weighted imaging might be that T1-weighted images contains sharper contrast around anatomical region borders, increasing the effect of slight variations in segmentation method. Our results do therefore not present a clear preference in the robustness of one MR modality over another.

Reproducibility of shape and first-order radiomic feature classes generally outperformed higher-order feature classes. These results differ somewhat from findings on hippocampal segmentation that found textural features to be the most reproducible.[53] However, single feature robustness could be application dependent, meaning that a feature that is found to be highly precise for a certain dataset and disease could have poor stability when assessed for another dataset or disease.[96] In both reproducibility and robustness, the GLSZM feature class stood out as the worst performing. On a feature level, robustness varies greatly as can be seen by the IQR in Fig. 2 and high variance in ICC in Supplementary Fig. S2.

Last, our results show that on an anatomical region level, radiomic feature reliability is relatively independent of anatomical region, apart from the CSF, choroid plexus, and inferior lateral ventricle. In these anatomically less relevant regions, low T1-intensity and low FA values are likely at the root of low reliability.

3.3. Robustness of radiomics-based prediction

In our work, previous findings in literature and clear robustness findings allowed for manual radiomic feature selection. As opposed to data-driven selection, manual selection allows for simplifying of the comparison of the different pipelines. Feature selection inherently means that some relations might have been lost and are not investigated in this work. However, model optimization and proving generalization of a predictive method on the topic of sleep deprivation is not within the scope of this work. Additionally, explanatory analysis of the contribution of radiomic features to this sleep deprivation classification lies outside of the scope of this work.

Despite the slight imbalance in the sleep deprivation labels (22/20) incurred by exclusion of subjects, accuracy still conveys the model performance in an intuitive way. Classification performance was significantly higher for the use of radiomic features derived from segmentation methods by Med-DeepBrain, and FastSurfer. Our work presents an independent comparison of methods that includes more segmentation methods than has previously been presented in literature. It is interesting to note that both segmentation methods with highest classification performance are not atlas-based, but use deep learning instead, which illustrates potential advantages of deep learning methods in segmentation. Future research should confirm that these methods perform equally well or better in other applications.

3.4. Study limitations

Main limitations of the study come down to the application sensitive nature of radiomics. First of all, with the data acquisition: Our study is limited to the two modalities used and shows differences in robustness metrics. Our results regarding robustness may not necessarily generalize to other modalities such as fMRI or ASL. Second, results regarding the differences in predictive performance of different segmentations methods might not generalize to other psychoradiological applications.

The dataset was relatively homogeneous and small in size. The population consisted of relatively young healthy individuals, which improves segmentation quality and potentially lowers disagreement among segmentation methods. Due to the dataset size, the size of the test set was limited which potentially increases variation in the cross-validation performance results. Since we did not stratify for labels in our partitioning scheme, some variation could also be attributed to slight imbalances due to the small sample size of our dataset.

Whole-brain segmentation only included 32 subcortical anatomical regions. We chose to limit the scope of this research to subcortical anatomical regions for two reasons: Cortical parcellation not available for SAMSEG, and cortical labeling definitions were inconsistent between VUNO Med-DeepBrain and FreeSurfer ASEG and FastSurfer methods.

3.5. Conclusion

Our work shows that small changes in segmentations due to a variation in image segmentation method affects radiomic features and subsequent predictive modeling when using these features. The robustness of these features is largely independent of the anatomical region, and lower-order radiomic features are generally more robust. Noteworthy, modern deep learning-based segmentation methods resulted in radiomic features that more accurately distinguishing sleep-deprived cases from controls. Our study suggests that methodological differences in fully automatic segmentation are of importance in radiomic feature-based cross-study comparison.

4. Material And Methods

4.1 Dataset

A scan-rescan randomized case-control MRI neuro-imaging dataset of healthy adults (n = 46, age 26 ± 7 years; 29 women) was used, as described in previous studies.[67–70] Twenty-three subjects were randomly assigned to either a night of sleep deprivation, or a normal sleep wake cycle (SWC). T1-weighted and DTI scans were acquired at four time points (TPs) over two days: TP1, around 9 a.m. after a night of normal sleep in their own home; TP2, around 8 p.m. approximately 11 hours after T1; TP3 approximately 23 hours after TP1; And finally, TP4, in the afternoon of the second day around 4 p.m. Participants in the normal SWC group went home to sleep between TP2 and 3, while those in the sleep deprivation group stayed at the hospital. This data collection was approved by the Regional Committee for Medical and Health Research Ethics, South-Eastern Norway (REK Sør-Øst, ref: 2017/2200) and conducted in line with the Declaration of Helsinki. All participants provided written informed consent. Data, code and documentation used in the study are available from the corresponding author upon reasonable request.

4.2 MRI acquisition and processing

Image acquisition consisting of T1-weighted and DTI and processing and was conducted in accordance with recommended standard brain segmentation protocols provided by the FreeSurfer group.[75]

T1-weighted brain images were scanned using a 3T Siemens Magnetom Prisma scanner (Siemens Healthcare, Erlangen, Germany) using a 32-channel head coil. The acquisition parameters[75] were as follows: repetition time (TR) = 2530ms, echo time (TE) = 3.5ms, flip angle = 7°. The voxel size was 1.0 × 1.0 × 1.0 mm and field-of-view (FOV) was 256 × 256 mm^2 (256 × 256 matrix) with 176 sagittal slices. Acquisition time was six minutes three seconds.

Preprocessing of T1 data consisted of motion correction, skull stripping and intensity normalization using the FreeSurfer image preprocessing pipeline (autorecon2).[76]

The DTI scan protocol has been described previously.[67] It consisted of a full-brain multi-shell Stejskal-Tanner pulsed mono-planar gradient scheme[77] with a single-shot spin-echo multiband-accelerated echo-planar imaging (EPI) readout module.[78] Seventy-six axial slices with b-values = [500-1000-2000-3000-4000](s/mm^2) and non-coplanar diffusion-sensitized gradient directions were acquired with the corresponding numbers of gradient directions n_{dir} = [12–30–40–50–60]. The following parameters were applied: TR = 2450ms, TE = 85ms, flip-angle = 78°. The voxel size = 2.0 × 2.0 × 2.0 mm, FOV = 212 x 212 mm^2 (106 × 106 matrix), slice thickness = 2 mm, and multiband acceleration factor = 4. Acquisition time was eight minutes and 21 seconds. In addition, five non-diffusion-weighted image sets (b = 0) of opposite phase-encode direction –but otherwise identical imaging parameters– were acquired for correction of susceptibility distortions. Acquisition time was 31 seconds.

Each DTI volume was affinely registered to the average non-diffusion weighted volume using the FMRIB's Linear Image Registration Tool (FLIRT)[79], correcting for intra-scan subject motion and eddy-current distortions. After non-brain tissue was removed[80], voxel-wise eigenvalues and eigenvectors were extracted from the estimated diffusion tensor and fractional anisotropy (FA) was calculated. The fractional anisotropy (FA) map was co-registered to the structural scan using the statistical parametric mapping (SPM) toolbox for Matlab 2019b (The MathWorks, Natick, Massachusetts). We manually checked for co-registration errors.

4.3 Segmentation and Segmentation agreement

Four whole-brain segmentation methods were selected: FreeSurfer Automatic Segmentation (ASEG)[81] version 7.1.0, FreeSurfer Sequence Adaptive Multimodal Segmentation (SAMSEG)[65,66], FastSurfer[55] and VUNO Med-DeepBrain (VUNO Inc., Seoul, South Korea) version 1.0.1. A description of each of these methods is provided in Supplementary Note S3. The selection of segmentation methods used was based on three considerations: First, the selection was limited to methods producing the same anatomical labeling. Second, methods were required to use the same modality as source, being T1-weighted MRI. And last, the different methods represent a mix of underlying methodologies. This mix consists of commonly used conventional and recently introduced methods that aim to outperform conventional methods through deep learning.

We generated segmentation labels from T1-weighted images and semantically matched these across segmentation methods, excluding areas that were not available for all methods. All segmentations were checked manually for all subjects to exclude potential failures. Segmentation agreement was determined by computing the Dice-coefficient for each segmentation method pair, for each anatomical area.[82] We interpreted Dice-coefficients using the same range of strength of agreement as for the Kappa coefficient and computed interquartile ranges (IQR) to aid interpretation.[83]

4.4 Radiomic feature extraction reliability analysis

We used seven classes to subdivide radiomic features, roughly in order of complexity: Shape-based features[84], first-order features[84], gray level co-occurrence matrices (GLCM)[85], gray level dependence matrices (GLDM)[86], gray level run length matrices (GLRLM)[87], gray level size zone matrices (GLSZM)[88], and neighboring gray tone difference matrices (NGTDM). A description of these feature classes can be found in Supplementary Note S4.

We computed radiomic features for each anatomical area, in each scan modality and at each TP using PyRadiomics[89] (v.3.0.1) implemented in Python (v.3.8.4). Geometry tolerance was set to 10^{-3} mm. Feature definitions are in compliance with Imaging Biomarker Standardization Initiative (IBSI)[44] and are described extensively at the proprietary repository.[90]

Next, we computed radiomic feature reliability for each anatomical area and each radiomic feature. Reliability was assessed for two measurements: First, scan-rescan reproducibility was calculated by computing the two-way mixed intra-class correlation (ICC)[91] between TP1 and TP2, thus before any effects of sleep deprivation. Second, robustness to segmentation method variation was computed similarly using ICC among all four segmentation methods. ICC was implemented using the Pingouin (v. 0.3.10) for Python. Resulting values were averaged over all subjects. Throughout this work, averaging of coefficients was performed using Fisher's z-transformation.[92] At first, computation of radiomic feature reliability produces a comprehensive overview of reliability of each feature for each MR modality. Second, reliability metrics were averaged per anatomical region and radiomic feature class. Anatomical regions with failing segmentation agreement were excluded to avoid including segmentation errors or regions with different semantical definitions that were not previously excluded in the matching of segmentation labels from affecting radiomic feature properties. Failure was defined as a Dice-coefficient below 0.5.

4.5 Radiomics-based classification and statistical analysis

To investigate the effect of the segmentation method variation on the discriminative power of radiomic features, we used a binary classifier to separate sleep-deprived subjects from controls. This classifier was trained on the radiomic features produced by each of the four segmentation methods.

Dimensionality of the data was reduced in three steps. First, a selection in modalities and radiomic features was made based on a combination of literary findings on effects of sleep deprivation and two

classes with highest radiomic feature reproducibility. Second, anatomical regions with failing agreement, as described in the previous section, were excluded. Left-right hemisphere values were not averaged, such that potential asymmetric properties remained. Last, to better express the temporal relationship in the data while reducing dimensionality, the values at the four time points were used to compute a first-order polynomial. Only the coefficient-over-time of this polynomial was used for prediction.

A neural network was trained on the remaining radiomic features. The input of the network consisted of a one-dimensional vector of radiomic features, and sleep deprivation as binary class label. The network consisted of three blocks, each consisting of a batch normalization layer, a rectified linear unit (ReLU) activation layer[93], linear layer and a dropout regularization layer, in that order. These blocks were followed by a sigmoid layer combined with a binary cross-entropy (BCE) loss.

The network was implemented in PyTorch version 1.7.1, on a single Nvidia GeForce RTC 2080 SUPER (Nvidia Corporation, Delaware, California) graphics processing unit (GPU) with CUDA version 10.2[94]. Adam optimization with L1-regularization, default weight initialization, and a constant learning rate of 10^{-3} were used for training and a batch size of 16. The regularization factor was set to 10^{-3} and dropout probability to 0.3. Training was performed until loss on the validation set refrained from decreasing. After training the parameters were reinstated to the state with lowest validation loss for testing.

A complete data approach, excluding subjects with missing scans, was followed. This is required to allow for the identical computation of the temporal effect using a polynomial. Data was divided into a training partition of 70%, validation partition of 20% and test partition of 10% of the included samples with which 10-fold cross-validation was performed. Each fold was initialized differently, but for each method the same sequence of randomization seed was used with the same partitioning to ensure comparability of the performance of the model of each fold.

Shapiro-Wilk test was used to test normality of the paired BCE-loss performance over all folds. Friedman test implemented in SciPy version 1.6.2 was used to test the hypothesis that model performance expressed as BCE-loss did not differ among methods used.[95]

Declarations

Acknowledgements

This work was supported by the Eurostars funding program (Reference number 113351) and research grants from the Norwegian South-East Health Authorities (reference numbers 2018077 and 2017090).

Author contributions

MGP, HM and MWAC conceived and designed the analysis. IG and AB collected the data. MGP performed the analysis and wrote the paper. HGR and LR contributed to the clinical relevance of the project. All authors contributed to and oversaw the content of the paper.

Data Availability Statement

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary material. The documented code base is available on GitHub (github.com/DEPREDICT/SLEEEP). Raw MRI data to support the findings of this study are available from the corresponding author, upon reasonable request

Competing Interests statement: M.W.A. Caan and H.A. Marquering are shareholders of Nico-lab International Ltd.

References

1. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. Radiology **278**, 563–577 (2016).
2. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer **48**, 441–446 (2012).
3. Kumar, V. *et al.* Radiomics: the process and the challenges. Magn. Reson. Imaging **30**, 1234–1248 (2012).
4. Avanzo, M., Stancanello, J. & El Naqa, I. Beyond imaging: The promise of radiomics. Phys. Medica **38**, 122–139 (2017).
5. Lui, S., Zhou, X. J., Sweeney, J. A. & Gong, Q. Psychoradiology: The Frontier of Neuroimaging in Psychiatry. *Radiology* **281**, 357–372 (2016).
6. Li, F., Wu, D., Lui, S., Gong, Q. & Sweeney, J. A. Clinical Strategies and Technical Challenges in Psychoradiology. Neuroimaging Clin. N. Am. **30**, 1–13 (2020).
7. Huang, X., Gong, Q., Sweeney, J. A. & Biswal, B. B. Progress in psychoradiology, the clinical application of psychiatric neuroimaging. Br. J. Radiol. **92**, 20181000 (2019).
8. Gong, J. *et al.* Predicting response to electroconvulsive therapy combined with antipsychotics in schizophrenia using multi-parametric magnetic resonance imaging. Schizophr. Res. **216**, 262–271 (2020).
9. Cui, L.-B. *et al.* Disease Definition for Schizophrenia by Functional Connectivity Using Radiomics Strategy. Schizophr. Bull. **44**, 1053–1059 (2018).
10. Xi, Y.-B. *et al.* Neuroanatomical Features That Predict Response to Electroconvulsive Therapy Combined With Antipsychotics in Schizophrenia: A Magnetic Resonance Imaging Study Using Radiomics Strategy. Front. Psychiatry **11**, 456 (2020).
11. Park, Y. W. *et al.* Differentiating patients with schizophrenia from healthy controls by hippocampal subfields using radiomics. Schizophr. Res. **223**, 337–344 (2020).
12. Gong, Q., Lui, S. & Sweeney, J. A. A Selective Review of Cerebral Abnormalities in Patients With First-Episode Schizophrenia Before and After Treatment. Am. J. Psychiatry **173**, 232–243 (2016).

13. Sun, H. *et al.* Psychoradiologic Utility of MR Imaging for Diagnosis of Attention Deficit Hyperactivity Disorder: A Radiomics Analysis. Radiology **287**, 620–630 (2018).
14. Wang, Y. *et al.* Classification of Unmedicated Bipolar Disorder Using Whole-Brain Functional Activity and Connectivity: A Radiomics Analysis. Cereb. Cortex **30**, 1117–1128 (2020).
15. Zhang, F.-F., Peng, W., Sweeney, J. A., Jia, Z.-Y. & Gong, Q.-Y. Brain structure alterations in depression: Psychoradiological evidence. CNS Neurosci. Ther. **24**, 994–1003 (2018).
16. Gong, Q. & He, Y. Depression, neuroimaging and connectomics: a selective overview. Biol. Psychiatry **77**, 223–235 (2015).
17. Antunes, J. *et al.* Radiomics Analysis on FLT-PET/MRI for Characterization of Early Treatment Response in Renal Cell Carcinoma: A Proof-of-Concept Study. Transl. Oncol. **9**, 155–162 (2016).
18. Chirra, P. *et al.* Empirical evaluation of cross-site reproducibility in radiomic features for characterizing prostate MRI. in *Medical Imaging 2018: Computer-Aided Diagnosis* (eds. Petrick, N. & Mori, K.) **10575**, 67–78 (SPIE, 2018).
19. Moradmand, H., Aghamiri, S. M. R. & Ghaderi, R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J. Appl. Clin. Med. Phys. **21**, 179–190 (2020).
20. Rathore, S. *et al.* Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. Sci. Rep. **8**, 1–12 (2018).
21. Zinn, P. O. *et al.* A Coclinical Radiogenomic Validation Study: Conserved Magnetic Resonance Radiomic Appearance of Periostin-Expressing Glioblastoma in Patients and Xenograft Models. Clin. cancer Res. an Off. J. Am. Assoc. Cancer Res. **24**, 6288–6299 (2018).
22. Bologna, M., Corino, V. & Mainardi, L. Technical Note: Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain. Med. Phys. **46**, 5116–5123 (2019).
23. Pandey, U., Saini, J., Kumar, M., Gupta, R. & Ingalhalikar, M. Normative Baseline for Radiomics in Brain MRI: Evaluating the Robustness, Regional Variations, and Reproducibility on FLAIR Images. J. Magn. Reson. Imaging **53**, 394–407 (2021).
24. Baeßler, B., Weiss, K. & Pinto Dos Santos, D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. Invest. Radiol. **54**, 221–228 (2019).
25. Lecler, A. *et al.* Combining Multiple Magnetic Resonance Imaging Sequences Provides Independent Reproducible Radiomics Features. Sci. Rep. **9**, 2068 (2019).
26. Balagurunathan, Y. *et al.* Test-retest reproducibility analysis of lung CT image features. J. Digit. Imaging **27**, 805–823 (2014).
27. Lubner, M. G., Smith, A. D., Sandrasegaran, K., Sahani, D. V & Pickhardt, P. J. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. *RadioGraphics* **37**, 1483–1503 (2017).

28. Mackin, D. *et al.* Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest. Radiol. **50**, 757–765 (2015).
29. Nyflot, M. J. *et al.* Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. J. Med. imaging (Bellingham, Wash.) **2**, 41002 (2015).
30. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci. Rep. **6**, 23428 (2016).
31. Mayerhoefer, M. E. *et al.* Effects of magnetic resonance image interpolation on the results of texture-based pattern classification: a phantom study. Invest. Radiol. **44**, 405–411 (2009).
32. Collewet, G., Strzelecki, M. & Mariette, F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn. Reson. Imaging **22**, 81–91 (2004).
33. Park, J. E., Park, S. Y., Kim, H. J. & Kim, H. S. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J. Radiol. **20**, 1124–1137 (2019).
34. Saha, A., Harowicz, M. R. & Mazurowski, M. A. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. Med. Phys. **45**, 3076–3085 (2018).
35. Cattell, R., Chen, S. & Huang, C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis. Comput. Ind. Biomed. art **2**, 19 (2019).
36. Lambin, P. *et al.* Radiomics: The bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. **14**, 749–762 (2017).
37. Park, J. E. & Kim, H. S. Radiomics as a Quantitative Imaging Biomarker: Practical Considerations and the Current Standpoint in Neuro-oncologic Studies. *Nucl. Med. Mol. Imaging (*2010*).* **52**, 99–108 (2018).
38. Pavic, M. *et al.* Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol. (Madr). **57**, 1070–1074 (2018).
39. Balagurunathan, Y. *et al.* Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Transl. Oncol. **7**, 72–87 (2014).
40. Parmar, C. *et al.* Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One **9**, e102107 (2014).
41. Ai, Y., Zhu, H., Xie, C. & Jin, X. Radiomics in cervical cancer: Current applications and future potential. Crit. Rev. Oncol. Hematol. **152**, 102985 (2020).
42. Fan, Y., Feng, M. & Wang, R. Application of Radiomics in Central Nervous System Diseases: a Systematic literature review. Clin. Neurol. Neurosurg. **187**, 105565 (2019).
43. Xue, C. *et al.* Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review. Quant. Imaging Med. Surg. **11**, 4431–4460 (2021).
44. Zwanenburg, A. *et al.* The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology **295**, 328–338 (2020).

45. Carré, A. *et al.* Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. Sci. Rep. **10**, 12340 (2020).
46. Ikeda, D. M. *et al.* Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies. J. Magn. Reson. Imaging **13**, 889–895 (2001).
47. Grimm, L. J. *et al.* Relationships Between MRI Breast Imaging-Reporting and Data System (BI-RADS) Lexicon Descriptors and Breast Cancer Molecular Subtypes: Internal Enhancement is Associated with Luminal B Subtype. Breast J. **23**, 579–582 (2017).
48. Wengert, G. J. *et al.* Inter- and intra-observer agreement of BI-RADS-based subjective visual estimation of amount of fibroglandular breast tissue with magnetic resonance imaging: comparison to automated quantitative assessment. Eur. Radiol. **26**, 3917–3922 (2016).
49. El Khoury, M. *et al.* Breast imaging reporting and data system (BI-RADS) lexicon for breast MRI: interobserver variability in the description and assignment of BI-RADS category. Eur. J. Radiol. **84**, 71–76 (2015).
50. Henderson, S. *et al.* Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer. Eur. Radiol. **27**, 4602–4611 (2017).
51. Saha, A. *et al.* Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. Med. Phys. **43**, 4558 (2016).
52. Park, J. E. *et al.* Diffusion and perfusion MRI radiomics obtained from deep learning segmentation provides reproducible and comparable diagnostic model to human in post-treatment glioblastoma. Eur. Radiol. (2020). doi:10.1007/s00330-020-07414-3
53. Li, Z., Duan, H., Zhao, K. & Ding, Y. Stability of MRI Radiomics Features of Hippocampus: An Integrated Analysis of Test-Retest and Inter-Observer Variability. IEEE Access **7**, 97106–97116 (2019).
54. Sederevi, D., Vidal-piñeiro, D., Sørensen, Ø., Leemput, K. Van & Eugenio, J. Reliability and sensitivity of two whole-brain segmentation approaches included in FreeSurfer – ASEG and SAMSEG. 1–28 (2020).
55. Henschel, L. *et al.* FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. Neuroimage **219**, 117012 (2020).
56. Suh, C. H. *et al.* Development and Validation of a Deep Learning{\textendash}Based Automatic Brain Segmentation and Classification Algorithm for Alzheimer Disease Using 3D T1-Weighted Volumetric Images. Am. J. Neuroradiol. (2020). doi:10.3174/ajnr.A6848
57. Bishop, C. A., Jenkinson, M., Andersson, J., Declerck, J. & Merhof, D. Novel Fast Marching for Automated Segmentation of the Hippocampus (FMASH): method and validation on clinical data. Neuroimage **55**, 1009–1019 (2011).
58. Doring, T. M. *et al.* Evaluation of hippocampal volume based on MR imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. J. Magn. Reson.

Imaging **33**, 565–572 (2011).

59. Merkel, B. *et al.* Semi-automated hippocampal segmentation in people with cognitive impairment using an age appropriate template for registration. J. Magn. Reson. Imaging **42**, 1631–1638 (2015).
60. Morey, R. A. *et al.* A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage **45**, 855–866 (2009).
61. Pardoe, H. R., Pell, G. S., Abbott, D. F. & Jackson, G. D. Hippocampal volume assessment in temporal lobe epilepsy: how good is automated segmentation? Epilepsia **50**, 2586–2592 (2009).
62. Mulder, E. R. *et al.* Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage **92**, 169–181 (2014).
63. Perlaki, G. *et al.* Comparison of accuracy between FSL's FIRST and Freesurfer for caudate nucleus and putamen segmentation. Sci. Rep. **7**, 2418 (2017).
64. Fischl, B. FreeSurfer. Neuroimage **62**, 774–781 (2012).
65. Puonti, O., Iglesias, J. E. & Van Leemput, K. Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. Neuroimage **143**, 235–249 (2016).
66. Puonti, O., Iglesias, J. E. & Van Leemput, K. Fast, Sequence Adaptive Parcellation of Brain MR Using Parametric Models. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (eds. Mori, K., Sakuma, I., Sato, Y., Barillot, C. & Navab, N.) 727–734 (Springer Berlin Heidelberg, 2013).
67. Voldsbekk, I. *et al.* Evidence for wakefulness-related changes to extracellular space in human brain white matter from diffusion-weighted MRI. Neuroimage **212**, 116682 (2020).
68. Elvsåshagen, T. *et al.* Widespread changes in white matter microstructure after a day of waking and sleep deprivation. PLoS One **10**, e0127351 (2015).
69. Elvsåshagen, T. *et al.* Cerebral blood flow changes after a day of wake, sleep, and sleep deprivation. Neuroimage **186**, 497–509 (2019).
70. Voldsbekk, I. *et al.* Sleep and sleep deprivation differentially alter white matter microstructure: A mixed model design utilising advanced diffusion modelling. Neuroimage **226**, 117540 (2021).
71. Elvsåshagen, T. *et al.* Evidence for cortical structural plasticity in humans after a day of waking and sleep deprivation. Neuroimage **156**, 214–223 (2017).
72. Bologna, M. *et al.* Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images. J. Digit. Imaging **31**, 879–894 (2018).
73. Zwanenburg, A. *et al.* Assessing robustness of radiomic features by image perturbation. Sci. Rep. **9**, 614 (2019).
74. Schwier, M. *et al.* Repeatability of Multiparametric Prostate MRI Radiomics Features. Sci. Rep. **9**, 9441 (2019).
75. Wiki, F. MORPHOMETRY PROTOCOLS. (2009). Available at: https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki?

action=AttachFile&do=get&target=FreeSurfer_Suggested_Morphometry_Protocols.pdf. (Accessed: 28th July 2021)

76. Dale, A., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. Neuroimage **9**, 179–194 (1999).
77. Stejskal, E. O. & Tanner, J. E. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. J. Chem. Phys. **42**, 288–292 (1965).
78. Setsompop, K. *et al.* NeuroImage Improving diffusion MRI using simultaneous multi-slice echo planar imaging. Neuroimage **63**, 569–580 (2012).
79. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. Med. Image Anal. **5**, 143–156 (2001).
80. Smith, S. M. Fast robust automated brain extraction. Hum. Brain Mapp. **17**, 143–155 (2002).
81. Fischl, B. *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron **33**, 341–355 (2002).
82. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. Ecology **26**, 297–302 (1945).
83. Benhajali, Y. *et al.* A Standardized Protocol for Efficient and Reliable Quality Control of Brain Registration in Functional MRI Studies. Front. Neuroinform. **14**, 7 (2020).
84. Rizzo, S. *et al.* Radiomics: the facts and the challenges of image analysis. Eur. Radiol. Exp. **2**, 36 (2018).
85. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural Features for Image Classification. IEEE Trans. Syst. Man. Cybern. **SMC-3**, 610–621 (1973).
86. Sun, C. & Wee, W. Neighboring gray level dependence matrix for texture classification. Comput. Graph. Image Process. **20**, 297 (1982).
87. Galloway, M. M. Texture analysis using gray level run lengths. Comput. Graph. Image Process. **4**, 172–179 (1975).
88. Thibault, G. *et al.* Shape and texture indexes - Application to cell nuclei classification. Int. J. Pattern Recognit. Artif. Intell. **27**, 1357002 (2013).
89. van Griethuysen, J. J. M. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. **77**, e104–e107 (2017).
90. van Griethuysen, J. J. M. Radiomic Features. *6 May* 2019 (2019). Available at: https://github.com/AIM-Harvard/pyradiomics/blob/master/docs/index.rst. (Accessed: 26th January 2021)
91. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. **15**, 155–163 (2016).
92. Silver, N. C. & Dunlap, W. P. Averaging correlation coefficients: should Fisher's z transformation be used? J. Appl. Psychol. **72**, 146 (1987).
93. Agarap, A. F. Deep learning using rectified linear units (relu). *arXiv Prepr. arXiv1803.08375* (2018).

94. NVIDIA, Vingelmann, P. & Fitzek, F. H. P. CUDA, release: 10.2.89. (2020).
95. Virtanen, P. *et al.* {SciPy} 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods **17**, 261–272 (2020).
96. Cattell, R., Chen, S. & Huang, C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis. Comput. Ind. Biomed. art **2**, 19 (2019).
97. Van Leemput, K. Encoding Probabilistic Brain Atlases Using Bayesian Inference. IEEE Trans. Med. Imaging **28**, 822–837 (2009).
98. Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989).

Table

Table 1 is available in the Supplemental Files section.

Figures

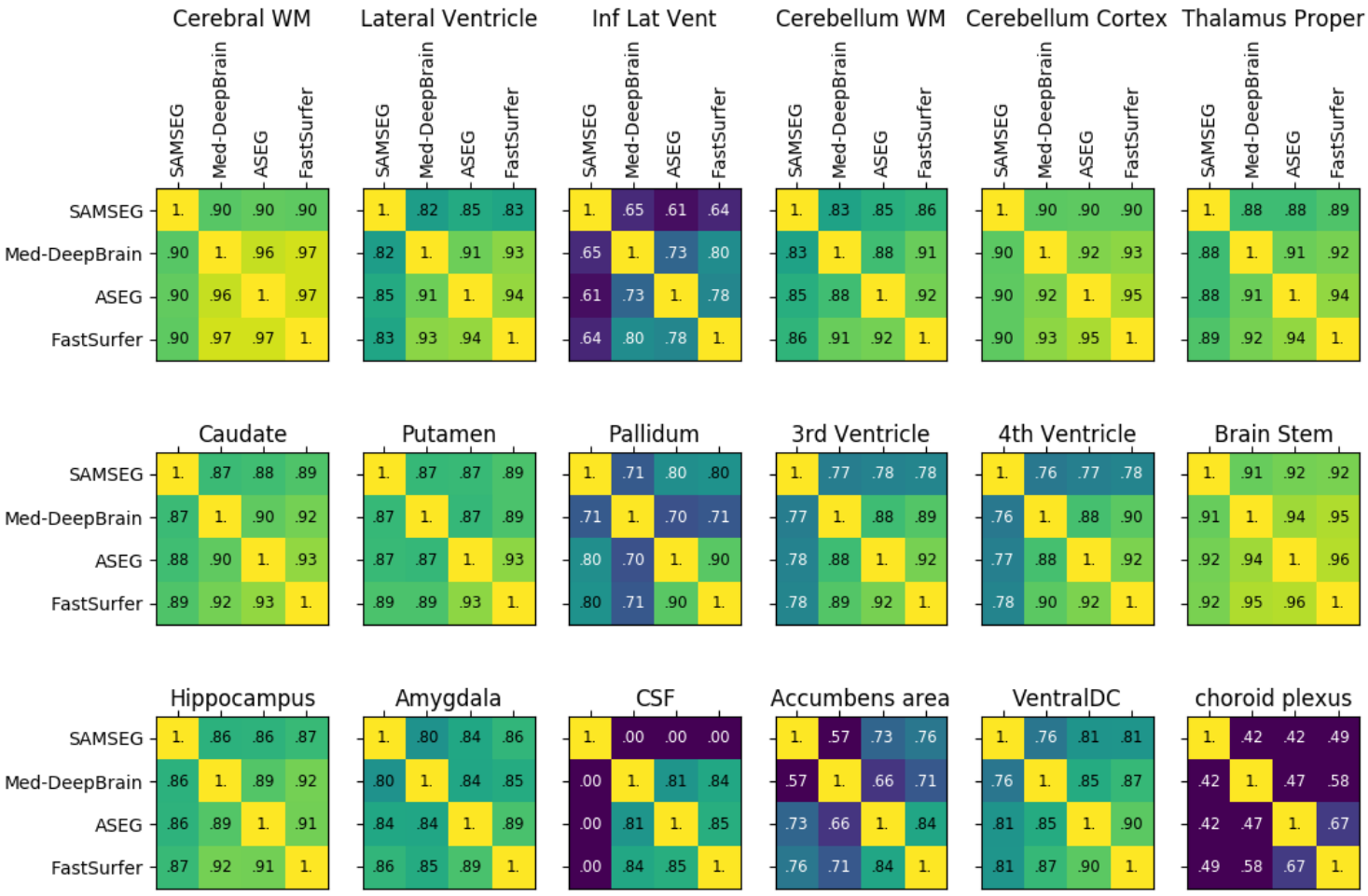

Figure 1

Pairwise segmentation agreement matrix. Dice-coefficients between each pair of segmentation methods for each subcortical area. Left-Right averages are shown where possible. Clarification of abbreviations: WM: white matter, DC: diencephalon, CSF: Cerebral Spinal Fluid, Inf Lat Vent: Inferior lateral ventricle

Figure 2

Reproducibility and robustness for each class of radiomic features. For an explanation of abbreviations see Appendix B: Description of radiomic feature classes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Radiomicsrobustnesstosegmentationvariationsupplement.pdf
- tab1.png