2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9433755

REGION SPECIFIC AUTOMATIC QUALITY ASSURANCE FOR MRI-DERIVED CORTICAL SEGMENTATIONS

Shruti Gadewar, Alyssa H. Zhu, Sophia I. Thomopoulos, Zhuocheng Li,
Iyad Ba Gari, Piyush Maiti, Paul M. Thompson, Neda Jahanshad

Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute,
Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

ABSTRACT

Quality control (QC) is a vital step for all scientific data analyses and is critically important in the biomedical sciences. Image segmentation is a common task in medical image analysis, and automated tools to segment many regions from human brain MRIs are now well established. However, these methods do not always give anatomically correct labels. Traditional methods for QC tend to reject statistical outliers, which may not necessarily be inaccurate. Here, we make use of a large database of over 12,000 brain images that contain 68 parcellations of the human cortex, each of which was assessed for anatomical accuracy by a human rater. We trained three machine learning models to determine if a region was anatomically accurate (as 'pass', or 'fail') and tested the performance on an independent dataset. We found good performance for the majority of labeled regions. This work will facilitate more anatomically accurate large-scale multi-site research.

Index Terms— Quality control, cortical parcellation, machine learning, Light Gradient Boost (LGBM), accuracy, precision, recall, F1-Score

1. INTRODUCTION

Quality control (QC) is an important step for all scientific analyses and is especially crucial in neuroimaging. Artifacts, noise, and poor image contrast are common in MRI-based imaging. These can often have detrimental effects on the quality of feature extraction, particularly for automatic image processing methods. For example, FreeSurfer (FS; Fischl 2012) is a widely used tool for brain tissue segmentation and produces individual cortical and subcortical parcellations according to specific atlases. Even so, severe atrophy or scanning artifacts may result in under- or over-estimation of cortical measurements. However, QC protocols often require a manual process with human interaction, which is time consuming, typically involving hours of training to learn. As large MRI studies become more common, manual QC becomes less feasible, and may become prohibitive.

Supervised machine learning algorithms can be used to learn informative patterns to classify data segmentations as accurate or not. A few supervised machine learning models have been introduced to automate quality control. One data-driven approach is Qoala-T [1], which compares various methods such as support vector machines, random forests, and gradient boosting methods, to automate overall quality control of FS segmentations of T1-weighted images but not QC for specific ROIs.

'Visual QC' provides an interface to visualize FS cortical parcellations and detect scanner-related variability [2]. Outlier-based methods have also been developed for quality assurance, such as LONI QC [3], MRIQC [4] and MindControl [5].

2. METHODS

2.1. Datasets

The UK Biobank (UKB) [6], Philadelphia Neurodevelopmental Cohort (PNC) [7], Parkinson's Progression Markers Initiative (PPMI) [8] and Alzheimer's Disease Neuroimaging Initiative (ADNI) [9] were analyzed in this work and are summarized in Table 1.

Table 1. Site specific details.

Data	Age	Control/ Cases Count	No. of Site	Field Strength and Manufacturer	Train/ Test count
UKB (4945 males)	62.5 +- 7.04	10350 population	4	3T Siemens	7245/3105
PPMI (373 males)	61.5 +- 10	161/486	24	1.5T and 3T Siemens, GE, Philips	453/194
PNC	14 +- 2.3	926 population	1	3T Siemens	649/277
ADN1 (476 males)	75 +- 7.4	228/595	50	1.5T from GE, Siemens, Philips	0/823

2.2. FreeSurfer (FS) Processing and Quality Assurance

FS v5.3's 'recon-all' function was run on all T1-weighted MRI scans from the datasets listed in Table 1. FS processing includes reconstruction of both the gray/white and pial surfaces, measuring cortical thickness, surface area and folding. We pilot our region-specific method for quality assurance with a coarse parcellation of the cortex, based on

978-1-6654-1246-9/21/$31.00 ©2021 IEEE

the Desikan-Killiany (DK) [10] set of 34 parcellations per hemisphere.

The ENIGMA Consortium QC protocol was used to manually assess whether regional parcellations were correct, according to anatomical boundaries. This protocol involves visual inspection of both 2D internal and external (surface) images [11]. The internal images provide insight into accurate thickness estimates (over-/under-estimation of the gray matter (GM) ribbon), while the surface maps help to ensure that the regions are labeled with the appropriate labels, as defined by [10]. Segmentation fails are tabulated in a spreadsheet.

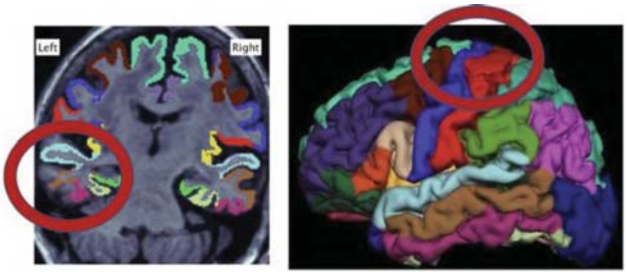


Fig. 1. Left) 2D coronal view that shows the FS cortical labels did not fully capture the cortical GM for one subject's left temporal lobe. The segmentations cut through the tissue. Right) We show the external view for QC. Parcellations include extracerebral tissue, contributing to a flat appearance.

2.3. Feature Set

All available cortical measures were extracted for each of the parcellated regions from the Desikan-Killiany atlas to create the feature set. This set comprises of regional measures including volume (grayvol), surface area (surfavg), thickness (thickstd and thickavg), number of vertices (numvert), and curvature measures like folding index (foldind), intrinsic curvature index (curvind), integrated rectified Gaussian curvature (gauscurv), mean curvature index (meancurv) and global measures including intracranial volume (ICV), left and right average thickness (LThickness, RThickness), left and right surface area average (LSurfavg, RSurfavg). The manual QC results were used as labels for model training. QC statuses were combined bilaterally per region of interest (ROI), resulting in a single label. If an ROI failed in the left and/or right hemisphere, the label was set to '1', otherwise 0.

2.4. Desikan-Killiany -Auto-QC Models

We compared three models using balanced accuracy (average of recall obtained for every class), precision (correct predicted "pass" / total predicted "pass"), recall (correct predicted "pass" / true "pass") and specificity (correct predicted "fail" /true "fail") as performance metrics. As the data is unbalanced, and most of the QC segmentations are correct and only a few of them are labeled as fails, we set the parameter "is_unbalanced" to 'True' in the classifier for all our approaches to account for sampling bias.

In model 1, we used XGboost Network, an optimized distributed gradient boosting library [12]. Gradient boosting is a technique where new models are created that predict the residuals or errors of prior models and then add them together to make a final prediction. It uses a gradient descent algorithm to minimize the loss while adding new models.

In model 2, we used the light gradient boost framework (LGBM) with gradient boosted trees as the base estimator. LGBM framework uses tree-based algorithms as base estimators [13]. Advantages of this model include lower memory usage, parallel and distributed learning, handling large-scale datasets and high efficiency. This algorithm achieves higher accuracy due to a leaf-wise tree growth approach, as opposed to other tree-based algorithms that use level-wise growth. LGBM grows trees leaf-wise and chooses the leaf with maximum delta loss to grow, Log loss was used here.

In model 3, we again used a LGBM but with a random forest as the base estimator. We set out 30% of the data from all cohorts to compare the predictions and performance metrics from all three models.

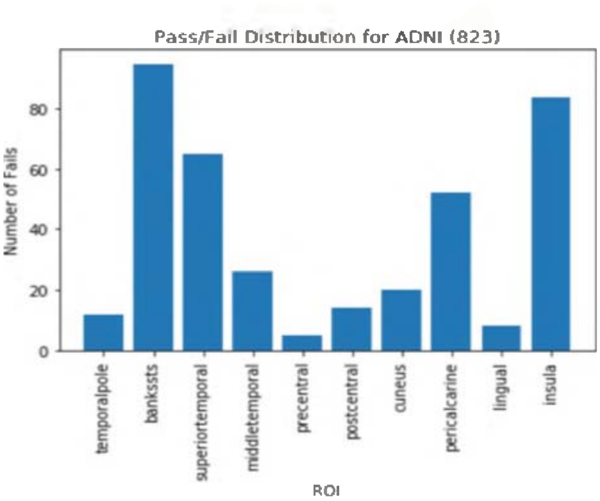


Fig. 2. Number of fails for a subset of ROIs that failed manual QC in the ADNI dataset (subject count: 823).

2.5. Training, Validation and Testing Samples

The UKB dataset was split into training (70%) and test (30%) sets; the training set was further split into an 80% training and 20% validation set to find the optimal parameters including the learning rate, maximum depth of the trees, and bagging fraction (number of samples included at every split in a tree) for model training. The data splitting was performed per ROI to keep the pass/fail ratio of the labels the same for training and testing. 5-fold cross validation was performed while training to choose best model parameters for models 2 and 3.

To train our models, we used 70% of the UKB PPMI, and PNC cohorts, while the remaining 30% from each were left out for testing (Table 1). The trained model was also tested on ADNI1 baseline data.

3. RESULTS

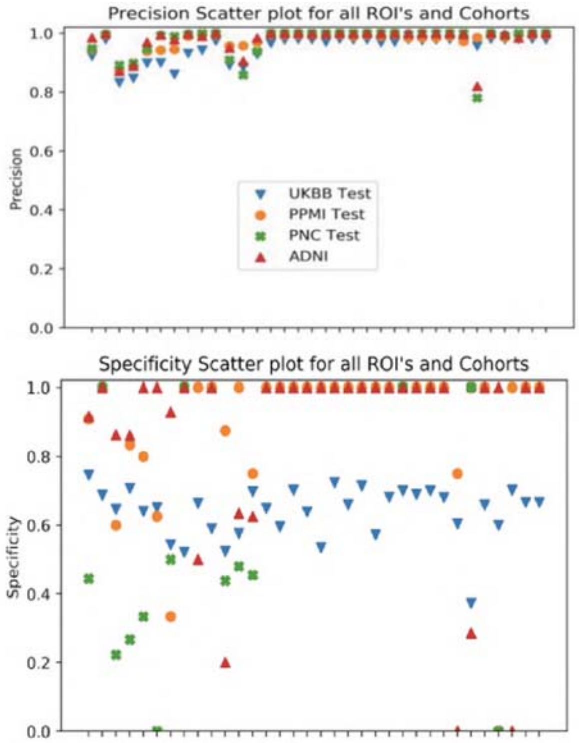
3.1. Performance Metrics

Table 2 shows the results for model 1 (XGboost), model 2 (LGBM with gradient boosted trees) and model 3 (LGBM with random forest). Model 1 and 2 perform worse on the task of predicting the "fail" labels, compared to model 3, for most ROIs. Model 2's performance for predicting the "passes" is consistently high across all the ROIs but low for predicting the "fails". This model passes almost every ROIs, hence the low specificity.

Model 3's true fail rate is lower for the "superior parietal", "cuneus" and "insula" in the ADNI test set, but performance for other ROIs is acceptable. The model fails more regions compared to model 1 and 2, which leads to low recall for all the ROIs in PPMI and ADNI test sets, both with a higher frequency of patient groups.

3.2. Feature Importance

While all ROI features were included for the QC of an individual region, as expected, the measures for the particular region itself, or the self-ROI measures were always among the top 10 features for each ROI classification. The most important self-features included: folding index (foldind), intrinsic curvature index (curvind), integrated rectified Gaussian curvature (gauscurv), thickness standard deviation (thickstd) and region-specific GM volume (gray vol) in decreasing order of importance.



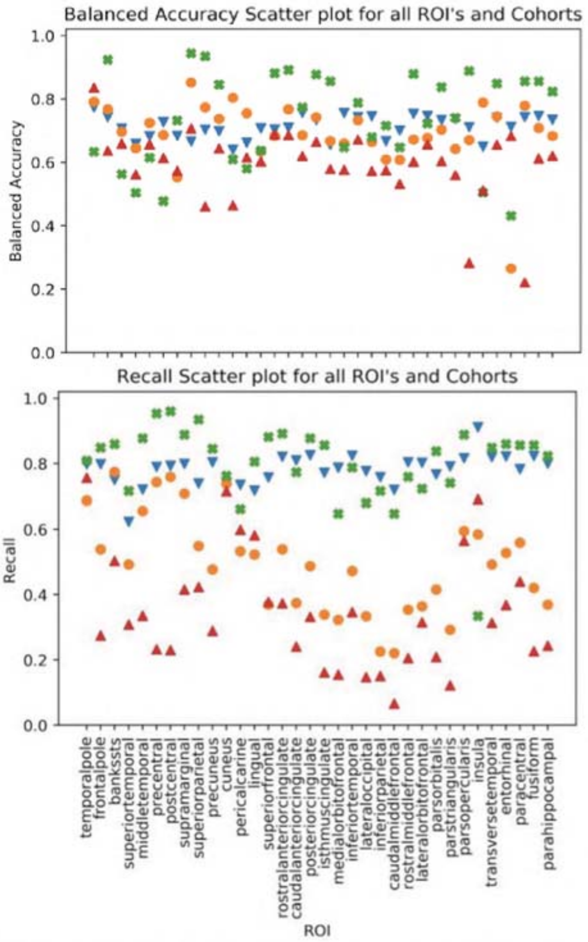


Fig. 3. Precision, Specificity, Balanced Accuracy and Recall scatter plot for all cohorts using model 3.

Table 2. Performance metrics for test datasets using model 1 (XGBoost), model 2 (LGBM with gradient boosted decision trees) and model 3 (LGBM with random forest).

DataSet FS v5.3	Control/Patient count	Precision m1,m2,m3	Specificity m1,m2,m3	Balanced Acc m1,m2,m3	Recall m1,m2,m3
UKB Test	3105 / 0	0.96 0.94 0.95	0.18 0.12 0.64	0.59 0.55 0.71	0.96 0.96 0.78
PPMI Test	47/147	0.97 0.97 0.98	0.03 0.04 0.90	0.51 0.52 0.69	0.98 0.98 0.50
PNC Test	277 / 0	0.97 0.97 0.97	0.02 0.05 0.51	0.80 0.80 0.74	0.98 0.98 0.80
ADNI 1	228 / 595	0.97 0.97 0.98	0.03 0.04 0.85	0.51 0.51 0.59	0.98 0.98 0.34

4. DISCUSSION AND CONCLUSIONS

We have compared QC results from three ML models and across all regions of interest, the LGBM with random forest outperformed both XGBoost and LGBM with gradient boosted trees. While cortical thickness and surface area were the features of biological interest, the most predictive regional metrics for QC were neither of those two, but instead related to folding and curvature, and the variability of thickness within the region. QC methods for cortical parcellations are increasingly important as several deep learning methods are emerging to emulate the cortical partitions produced by FS. FastSurfer [14], for example, uses deep learning to provide a full FS alternative for volumetric analysis (in under 1 minute) and surface-based thickness analysis (within only around 1 h runtime), that can be run on tens of thousands of brain MRI datasets in a fraction of the time required to run FS. Meanwhile, other research groups have developed tools to create alternative cortical parcellations (e.g., MUSE [15]), that calculate a consensus segmentation by fusion of anatomical labels from multiple atlases and registrations and obtain more consistent segmentations for some brain structures across scanners, compared to FS.While these tools will allow for fast parcellation of the cortex, their anatomical accuracy may be overlooked.

Our work provides a fast and reliable tool to ensure label accuracy. Our trained models will be available on GitHub.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was conducted retrospectively using publicly available, anonymized, de-identified human subject's data. Additional ethical approval was not required beyond that obtained in the original studies.

6. ACKNOWLEDGEMENTS

This research was supported in part by the National Institutes of Health (NIH) under grants R01AG059874, RF1AG051710, U01AG068057, and P41EB015922. No conflict of interest. This research was conducted using the UK Biobank Resource under Application Number 11559, Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data) and Philadelphia Neurodevelopmental Cohort (PNC) database.

7. REFERENCES

[1] Klapwijk, E. T., Kamp, F. van de, Meulen, M. van der, Peters, S., & Wierenga, L. M. "Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data." NeuroImage, (2019, January 8): pp. 323-329

[2] Raamana, P. R., Theyers, A., Selliah, T., Bhati, P., Arnott, S. R., Hassel, & Milev, R. (2020). Visual QC protocol for FreeSurfer cortical parcellations from anatomical MRI. BioRxiv, Oct. 2020

[3] Kim, Hosung et al. "The LONI QC System: A Semi-Automated, Web-Based and Freely-Available Environment for the Comprehensive Quality Control of Neuroimaging Data." Frontiers in Neuroinformatics, vol. 13, 60. 28 Aug. 2019, doi:10.3389/fninf.2019.00060

[4] Esteban, Oscar, et al. "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites." PloS one 12.9 (2017): e0184661.

[5] Keshavan, Anisha, et al. "Mindcontrol: A web application for brain segmentation quality control." NeuroImage 170 (2018): 365-372.

[6] Miller, Karla L., et al. "Multimodal population brain imaging in the UK Biobank prospective epidemiological study." Nature Neuroscience 19.11 (2016): pp. 1523-1536

[7] Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., ... & Mentch, F. D. "Neuroimaging of the Philadelphia Neurodevelopmental Cohort." NeuroImage 86 (2014): pp. 544-553

[8] Marek, K., et al. "The Parkinson Progression Marker Initiative (PPMI)." Progress in Neurobiology 95.4 (2011): pp. 629-635

[9] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C.R. Jack, W. Jagust, et al. "Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI)" Alzheimers Dement 1 (2005): pp. 55-66

[10] Alexander, B., Loh, W. Y., Matthews, L. G., Murray, A. L., Adamson, C., Beare, R., ... & Spittle, A. J. (2019). "Desikan-Killiany-Tourville atlas compatible Version of M-CRIB neonatal parcellated whole brain atlas: the M-CRIB 2.0." Frontiers in Neuroscience, pp.13-34

[11] Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., ... & Wright, M. J. "The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data." Brain Imaging and Behavior 8(2) (2014): pp. 153-182

[12] Chen, T., & Guestrin, C. (2016, August). XGboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIG-KDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794)

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "LGBM: A Highly Efficient Gradient Boosting Decision Tree." Advances in Neural Information Processing Systems 30 (NIPS 2017): pp. 3149-3157

[14] Henschel, Leonie, et al. "FastSurfer-A fast and accurate deep learning-based neuroimaging pipeline." NeuroImage (2020): 117012.

[15] Srinivasan, Dhivya, et al. "A comparison of FreeSurfer and multi-atlas MUSE for brain anatomy segmentation: Findings about size and age bias, and inter-scanner stability in multi-site aging studies." NeuroImage 223 (2020): 117248.