frontiers

Article Title

Steven Giavasis [1], **Sang Han Lee** [2], **Zarrar Shehzad** [1,2,3], **Qingyang Li** [1], **Yassine Benhajali** [4,5], **Chaogan Yan** [6], **Zhen Yang** [7], **Michael Milham** [1,8], **Pierre Bellec** [5], **R. Cameron Craddock** [1,2,*],

[1] *Center for the Developing Brain, Child Mind Institute, New York, NY, USA*
[2] *Computational Neuroimaging Laboratory, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA*
[3] *Department of Pscyhology, Yale University, New Haven, CT, USA*
[4] *Département d'anthropologie, Université de Montréal, Montréal, QC, Canada*
[5] *Centre de recherche de linstitut de gériatrie de Montréal, Montréal, QC, Canada*
[6] *Institute of Psychology, Chinese Academy of Science, Beijing, China*
[7] *Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA*
[8] *Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA*

Correspondence*:
R. Cameron Craddock
Computational Neuroimaging Laboratory, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, NY, 10962, USA, ccraddock@nki.rfmh.org

ABSTRACT

For full guidelines regarding your manuscript please refer to Author Guidelines or **Table ??** for a summary according to article type.

Keywords: Text Text Text Text Text Text Text Text

1 INTRODUCTION

It is well accepted that poor quality data interferes with the ability of neuroimaging analyses to uncover biological signal and distinguish meaningful from artifactual findings, but there is no clear guidance on how to differentiate good from bad data. A variety of different measures have been proposed, but there is no evidence supporting the primacy of one measure over another or on the ranges of values for differentiating good from bad data. As a result, researchers are required to rely on painstaking visual inspection to assess data quality. But this approach consumes a lot of time and resources, is subjective, and is susceptible to inter-rater and test-retest variability. Additionally, it is possible that some defects are too subtle to be fully appreciated by visual inspection, yet are strong enough to degrade the accuracy of data processing algorithms or bias analysis results. Further, it is difficult to visually assess the quality of data that has already been processed, such as that being shared through the Preprocessed Connectomes Project (PCP; http://preprocessedconnectomesproject.github.io/), the Human Connectome Project (HCP), and the Addiction Connectomes Preprocessing Iniatiative (ACPI). To begin to address this problem, the PCP has

assembled several of the quality metrics proposed in the literature to implement a Quality Assessment Protocol (QAP; http://preprocessedconnectomesproject.github.io/qualityassessmentprotocol).

The QAP is a open source software package implemented in python for the automated calculation of quality measures for functional and structural MRI data. The QAP software makes use of the Nipype () pipelining library to efficiently acheive high throughput processing on a variety of different high performance computing systems (). The quality of structural MRI data is assessed using contrast-to-noise ratio (CNR; Magnotta and Friedman, 2006), entropy focus criterion (EFC, Atkinson 1997), foreground-to-background energy ratio (FBER,), voxel smoothness (FWHM, Friedman 2008), percentage of artifact voxels (QI1, Mortamet 2009), and signal-to-noise ratio (SNR, Magnotta and Friedman (2006)). The QAP includes methods to assess both the spatial and temporal quality of fMRI data. Spatial quality is assessed using EFC, FBER, and FWHM, in addition to ghost-to-signal ratio (GSR). Temporal quality of functional data is assessed using the standardized root mean squared change in fMRI signal between volumes (DVARS; Nichols 2013), mean root mean square deviation (MeanFD, Jenkinson 2003), the percentage of voxels with MeanFD ¿ 0.2 (Percent FD; Power 2012), the temporal mean of AFNIs 3dTqual metric (1 minus the Spearman correlation between each fMRI volume and the median volume; Cox 1995) and the average fraction of outliers found in each volume using AFNIs 3dTout command.

Applying the QAP for differentiating good quality data from poor will require learning which of the measures are the most sensitive to problems in the data and the ranges of values for the measures that indicate poor data. The solutions to these questions are likely to vary based on the analyses at hand and finding them will likely require the ready availability of large scale hetereogenous datasets for which the QAP metrics have been calculated. To help with this goal, the QAP has been used to measure structural and temporal data quality on data from the Autism Brain Imaging Data Exchange (ABIDE; Di Martino 2013) and the Consortium for Reliability and Reproducibility (CoRR, Zuo 2014) and the results are being openly shared through the PCP. An initial analyses of the resulting values has been performed to evaluate their collinearity, correspondence to expert-assigned quality labels, and test-retest reliability.

2 METHODS

The Preprocessed Connectomes Project Quality Assessment Protocol is an open source toolkit of quality assessment measures implemented in python. Calculating the measures requires several standard preprocessing procedures such as tissue segmentation, image registration and mask generation that are accomplished using components from FSL(;) and AFNI (AFNI;). These tools are integrated with QAP specific python functions using the Nipype () pipelining library to automate processing of very large datasets in parallel on high performance computing systems such as multicore workstations and clusters that use Sun Grid Engine. The toolkit provides everything necessary to calculate the measures from scratch using raw imaging data but can also import data intermediaries (i.e. tissue segmentation maps) processed outside of the QAP pipeline. The software requires the images to be in the NIfTI file format and can handle a variety of different directory structures using an easy to construct configuration file. The software can be installed using standard python package installation tools (e.g. pip) and is available preinstalled on a free-to-use Amazon Machine Instance. Extensive documentation on installing and using the toolkit are available at its webpage ().

3 THE METRICS

The toobox includes a variety of different metrics for assessing spatial and temporal quality of data that have been proposed in the literature. The goal has been to make the toolbox comprehensive even though many of the measures may be highly correlated. Measures, such as signal-to-noise-fluctuation ratio (also known as temporal signal to noise ratio), that are only appropriate for phantom studies and have been excluded along with measures, such as QI2, that are overly complicated or computationally expensive with marginal sensitivity to quality (). When possible the amount of processing required for calculating the images has been minimized so that the measure focuses on the quality of the data rather than the quality of the algorithms used to perform the processing. But since some processing, such as segmentation, alignment, and masking, is unavoidable we recommend that QA measures be calculated using the same algorithms. QAPs current version only includes measures for structural and functional MRI data, measures for other imaging modalities such as diffusion MRI will be added in the future.

3.1 Measures of spatial quality

3.1.1 Contrast-to-Noise Ratio (CNR)

CNR can be defined in many different ways depending on the purpose of the images being collected. Since structural MRI data is most commonly used for morphometric measurements and calculating tissue specific maps for downstream processing, the QAP focuses on the contrast between white matter and grey matter. CNR should correlate with how well anatomical features can be discerned from the image and provides a measure of how easily the image can be segmented. It is sensitive to the imaging parameters used to acquire the data, as well as, the amount of thermal noise, artifacts, and head motion present in the image. Higher values for CNR are better. CNR is only calcuated for structural MRI data.

$$CNR = \frac{\overline{WM} - \overline{GM}}{\sigma_b} \tag{1}$$

CNR is the difference between the mean white matter signal and the mean gray matter signal divided by the standard deviation of the image background (Eqn. 1). The input structural MRI image is segmented into grey matter and white matter masks using FSLs FAST. Image background is defined by inverting a whole head mask that was defined by sequentially dilating and eroding the result of AFNIs 3dAutomask 4 times. The resulting GM and WM masks are applied to the input image to calculate the mean voxel intensity within each comparment, and the background mask is used to calculate the standard deviations of voxels outside of the head.

3.1.2 Entropy Focus Criterion (EFC)

EFC is calculated as the Shannon entropy of the voxel intensities. This value reaches its maximum when the number of voxels in the image are spread evenly over the different voxel intensities present in the image and will reach its minimum if the voxels are disproportionately focused on a single intensity value. Since the majority of voxels in a structural brain image should have zero intensity, which are typically viewed as black, this measure is often interpreted as the blackness of the image. Ghosting and head motion should reduce the amount of black in the image background, and hence should increase EFC (Atkinson 1997). The lower this number is, the better.

3.1.3 Foreground-to-Background Energy Ratio (FBER)

FBER is calculated by dividing the energy (normalized variance) of the image foreground by the energy of the background. In an ideal brain image all of the image energy should be contained in the foreground. Head motion, thermal noise, ghosting, and other artifacts will increase the energy in the images background and will reduce this value. Calculating FBER requires a head mask which is constructed by iterative dilating and eroding the output of 3dAutomask 4 times, backgroud is defined as the inversion of this mask. Higher values of FBER are better.

3.1.4 Full-Width Half Maximum (FWHM)

This spatial metric measures how smoothed the image is, which is essentially a measure of the degree of spatial correlation in the image data. The QAP pipeline uses the AFNI command 3dFWHMx to calculate this value. The lower this number is, the better.

3.1.5 Artifact Detection (Qi1)

For this measure, the proportion of voxels with intensity corrupted by artifacts is normalized by the number of voxels in the background (the air). The lower this number is, the better.

3.1.6 Signal-to-Noise Ratio (SNR)

Given the anatomical segmentation maps and the anatomical head mask, the mean of the gray matter signal is calculated and then divided by the standard deviation of the mean of the signal values within the background (air). The higher this number is, the better.

Like the spatial measures for anatomical data, the spatial quality measures for functional data include EFC, FBER, FWHM, and SNR. Unlike the spatial measures for anatomical data, however, these measures are run on the mean of the functional timeseries, which preserves the spatial features of the functional timeseries data. As mentioned above, in addition to these measures, there is another metric exclusive to the functional spatial metrics:

3.1.7 Ghost-to-Signal Ratio (GSR)

The GSR is the mean of the signal in the ghost of the image (artifacts appearing outside of the brain, caused by phase discontinuities in the phase-encoding direction) relative to the mean signal within the brain. The user must know the phase-encoding direction related to the scans being analyzed. The lower this number is, the better.

3.2 Measures of temporal quality

3.2.1 Standardized DVARS (Power 2012)

This measure is calculated by normalizing the spatial standard deviation of the temporal derivative of the data with the temporal standard deviation and temporal autocorrelation. The standardization, calculated by a script by Nichols [cite], provides a more absolute measure of DVARS which can be compared across subjects. The lower this value is, the better.

3.2.2 Outlier Detection

The AFNI tool 3dTout is used to find the mean fraction of outliers in each volume of the timeseries. The lower this value is, the better.

3.2.3 Global Correlation (GCOR)

The global correlation measure is the average correlation of every combination of voxels in the functional time series. The difference in GCOR between time series can help identify differences in the data that arise from motion or physiological noise.

3.2.4 Median Distance Index (Quality)

The AFNI tool 3dTqual is used to calculate a quality index describing the functional timeseries. This tool finds the volume of median value and then calculates the mean distance (Spearmans rho) between each volume and the median. The lower this value is, the better.

3.2.5 Mean Framewise Displacement (Jenkinsons Mean FD)

A measure of subject head motion, which compares the motion between the current and previous volumes. This is calculated by summing the absolute value of displacement changes in the x, y and z directions and rotational changes about those three axes. The rotational changes are given distance values based on the changes across the surface of a 80mm radius sphere. The lower this number is, the better.

3.2.6 Number of volumes with FD greater than 0.2mm (Num_FD)

This is the number of volumes in the timeseries whose Jenkinsons Mean FD exceeds 0.2. The lower this number is, the better.

3.2.7 Percent of volumes with FD greater than 0.2mm (Perc_FD)

This is the percent of total volumes in a timeseries whose Jenkinsons Mean FD exceeds 0.2. The lower this number is, the better.

3.3 The QAP Pipeline Python Toolbox

3.3.1 Software Description

The QAP measures pipelines were implemented in part using Nipype, an open-source neuroinformatics software project which allows the streamlining of neuroimaging processing pipelines [CITE]. As the metrics employed can be grouped by which type of data they are used to assess, a pipeline builder and runner was created for each of these groups: anatomical spatial measures (for anatomical/structural scans), functional spatial measures (for the spatial characteristics of a functional 4D timeseries), and functional temporal measures (for the characteristics of the timeseries themselves).

Each pipeline runner script accepts a subject list and a configuration file. These files are accepted as the Python YAML file format, which is a user-friendly file format allowing the user to list labeled options easily in a text editor. The subject list can contain file paths to either raw data scans, or already-preprocessed intermediary files. The configuration file contains a small collection of settings for the pipeline, such as how many processors to dedicate to the pipeline, or which template brain to use for steps like registration. The QAP software package includes scripts which help the user quickly generate these subject lists.

When raw data scans are provided to the pipelines, the necessary preprocessing steps required to complete the QAP metrics are automatically inserted into the pipeline and executed. Alternatively, if the user already has some or all of the preprocessing completed (for example, if the user preferred to complete anatomical-to-template registration their own way), these already generated intermediary files can be provided directly to the pipeline via the subject list, thereby passing all preprocessing steps up to that

specific step. In addition, the pipelines feature a ”warm restart” capability which allows the user to stop processing at any point, and later restart where it previously left off.

The pipelines are equipped to run the measures for multiple subjects in parallel. They can be run either through the command line interface or through Amazon Web Services' cloud computing infrastructure. Details on how to install, configure and run these scripts are provided in the QAP Python toolbox projects online documentation.

3.3.2 Calculation of Measures

The spatial quality measures for anatomical data listed above were calculated using the processing pipeline initiated using the qap_anatomical_spatial.py script of the QAP Python toolbox. They are calculated as follows:

was used to calculate spatial and temporal quality measures on the (now 1,101 structural and 1,163 functional) 1,113 structural and functional MRI datasets from the ABIDE dataset and the (now 3,112 structural and 4,611 functional scans) 3,357 structural and 5,094 functional scans from the CoRR dataset. For the ABIDE data, quality measures were compared to the quality scores determined from visual inspection by three expert raters to evaluate their predictive value. For both the ABIDE and CoRR datasets, the redundancy between quality measures was evaluated from their correlation matrix. Finally, the test-retest reliability of quality measures derived from CoRR was assessed using intraclass correlation.

3.4 The QAP Resource of Quality Measures

3.5 Methods for assessing the quality of ABIDE data

3.6 Statistical Analysis

4 RESULTS

Figure 1. Examples of measures and distributions for ABIDE (and CoRR) Figure 2. Correlogram of measures Figure 3. test retest of measures Figure 4. boxplots of most discriminative measures vs. hand assessments Figure 5. regression plots of most significant relationships with scanning parameters

Each of the measures showed a good bit of variability between imaging sites (see Figure 1 for an example plot showing standardized DVARS for ABIDE). Ranks calculated from the weighted average of standardized quality metrics indicated that CMU was the worst performing site and NYU was the best. QI1 and SNR were the best predictors of manually applied structural data quality scores, and EFC, FWHM, Percent FD, and GSR were all significant predictors of functional data quality (fig 2, p¡0.0001). A few of the measures are highly correlated (fig. 3) such as SNR, CNR and FBER, which measure very similar constructs, indicated that there is some room for reducing the set of measures. For the functional data, the test-retest reliability of several of the spatial measures of quality were very high (fig 4., EFC, FBER, GSR) reflecting their sensitivity to technical quality (i.e. MR system and parameters) whereas temporal measures were lower reflecting their sensitivity to physiological factors such as head motion. Similarly in the structural data, it appears that measures can be divided into those that are more sensitive to technical quality (EFC, FWHM) and those that favor physiological variation (CNR, QI1) based on test-retest reliability.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

The statement about the authors and contributors can be up to several sentences long, describing the tasks of individual authors referred to by their initials and should be included at the end of the manuscript before the References section.

ACKNOWLEDGMENTS

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Funding: Text Text Text Text Text Text Text Text.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

REFERENCES

Magnotta, V. A. and Friedman, L. (2006). Measurement of Signal-to-Noise and Contrast-to-Noise in the fBIRN Multicenter Imaging Study. *J Digit Imaging* 19, 140–147. [PubMed Central:PMC3045184] [DOI:10.1007/s10278-006-0264-x] [PubMed:16598643]

FIGURES

Figure 1. Enter the caption for your figure here. Repeat as necessary for each of your figures