The Preprocessed Connectomes Project Quality Assessment Protocol: a resource for measuring the quality of MRI data.

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

[1] *Center for the Developing Brain, Child Mind Institute, New York, NY, USA*
[2] *Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA*
[3] *Department of Psychology, Yale University, New Haven, CT, USA*
[4] *Department of Psychology, Stanford University, Stanford, CA, USA*
[5] *Département d'anthropologie, Université de Montréal, Montréal, QC, Canada*
[6] *Centre de recherche de linstitut de gériatrie de Montréal, Montréal, QC, Canada*
[7] *Institute of Psychology, Chinese Academy of Science, Beijing, China*
[8] *Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, 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 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 artefactual findings, but there is no clear guidance on how to differentiate good from bad data. A variety of measures for assessing data quality have been proposed (Magnotta and Friedman, 2006; Atkinson et al., 1997; Friedman et al., 2008; Mortamet et al., 2009; Power et al., 2012; Giannelli et al., 2010) add reference to new metric, but there is no consensus on the primacy of one measure over another or on the ranges of values for the measures that indicate poor quality 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://preprocessed-connectomes-project.github.io/`), the Human Connectome Project (HCP) (Van Essen and Ugurbil, 2012; Glasser et al., 2013), and the Addiction Connectomes Preprocessing Iniatiative (ACPI; `http://fcon_1000.projects.nitrc.org/indi/ACPI/html/`). 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://preprocessed-connectomes-project.github.io/quality-assessment-protocol`).

need to add new metricThe QAP is an open source software package implemented in Python for the automated calculation of quality measures for functional and structural MRI data. The QAP software combines functionality from the AFNI (Cox, 1996) neuroimaging toolkit with custom Python functions using the Nipype pipe-lining library (Gorgolewski et al., 2016b) to efficiently achieve 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 et al., 1997), foreground-to-background energy ratio (FBER), voxel smoothness (FWHM) (Friedman et al., 2008), percentage of artifact voxels (QI1) (Mortamet et al., 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) (Giannelli et al., 2010). Temporal quality of functional data is assessed using the standardized root mean squared change in fMRI signal between volumes (DVARS) (Power et al., 2012; Nichols, 2013), mean root mean square deviation (MeanRMSD) Jenkinson (1999), the temporal mean of AFNIs `3dTqual` metric (Cox, 1996), global correlation (GCOR) (Saad et al., 2013), and the average fraction of outliers found in each volume using AFNIs `3dTout` command (Cox, 1996).

Using QAP outputs for quantitatively (or automatically) assessing data quality will require learning which of the measures are the most sensitive to poor quality and the ranges of their values that indicate good data. The solutions to these questions are likely to vary based on the analyses at hand and finding them will require the ready availability of QAP metrics calculated on large scale heterogeneous datasets. 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 et al., 2014) and the Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 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

2.1 Quality Measures

The QAP toolbox includes a variety of metrics that have been proposed in the literature for measuring spatial and temporal aspects of structural and functional neuroimaging data. The goal has been to make the toolbox comprehensive even though many of the measures may be highly correlated. Measures from the literature that are explicitly defined for phantom data and are not appropriate for in vivo data, such as signal-to-noise-fluctuation ratio (also known as temporal signal to noise ratio) (Friedman and Glover, 2006), have been excluded along with measures such as noise distribution analysis (QI2) that are computationally expensive with marginal sensitivity to quality (Mortamet et al., 2009). QAP currently only includes measures for structural and functional MRI data, measures for other imaging modalities like diffusion MRI will be added in the future.

2.1.1 Measures of spatial quality

2.1.1.1 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, QAP focuses on the contrast between white matter and grey matter. CNR is therefore calculated as the difference between the mean white matter signal ($\overline{WM}$) and the mean gray matter signal ($\overline{GM}$) divided by the standard deviation of the image background (σ_b) (see Eqn. 1) (Magnotta and Friedman, 2006).

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

CNR should correspond to the ability to discern anatomical features 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. CNR is only calculated for structural MRI data, and the greater this value is, the better.

2.1.1.2 Entropy Focus Criterion (EFC)

EFC is the Shannon entropy present across voxel intensities, which is maximized when the voxel intensity histogram is spread evenly across all intensities and is minimized when voxels all have the same intensity. Head-motion induced image blurring and ghosting cause background voxels that would otherwise be zero to have a brighter intensity. The resulting spread of the voxel intensity histogram will result in greater entropy. Hence, this measure can be used to approximate the degree of motion-related artifacts in an image (Atkinson et al., 1997). EFC is computed using equation 2:

$$EFC = -\sum_{n=1}^{N} \frac{V_n}{V_{max}} \ln\left(\frac{V_n}{V_{max}}\right) \tag{2}$$

with N being the number of image voxels, and V_n being the intensity of the n^{th} voxel. V_{max} is proportional to the standard deviation of voxel intensities and is defined in equation 3.

$$V_{max} = \sqrt{\sum_{n=1}^{N} V_n^2} \tag{3}$$

The maximum value of EFC is determined by the number of voxels in the image (equation 2), therefore we divide EFC by the maximum to make the result comparable across images of different sizes.

$$EFC_{max} = \sqrt{N}\ln(\sqrt{N}) \tag{4}$$

EFC is calculated for both anatomical and functional data, and the closer to zero this number is, the better.

2.1.1.3 Foreground-to-Background Energy Ratio (FBER)

Foreground-to-Background Energy Ratio, or FBER, is the ratio between the energy (normalized variance) of the signal in voxels in the head and the energy of the signal in the background (the air). This provides a measure of how much of the signal is contributed by motion-related or scanner-related artifacts. In an ideal brain image all of the energy should be contained in the foreground. Head motion, thermal noise, ghosting, and other artifacts will increase the energy in an image's background and will reduce this value. FBER is calculated for both anatomical and functional data, and larger values of FBER are better.

$$FBER = \frac{\frac{1}{|F|}\sum_{f \in F} V_f^2}{\frac{1}{|B|}\sum_{b \in B} V_b^2} \tag{5}$$

where F is the set of voxels in foreground and B is the set of voxels in background.

2.1.1.4 Full-Width Half Maximum (FWHM)

FWHM is a measure of the the spatial smoothness - the degree of spatial correlation - in the imaging data. The spatial smoothness in a particular direction (x, y, or z) is estimated from the ratio of the variance of the first spatial difference image along the direction to the variance of the unperturbed image. The FWHM in each of the three directions are estimated and then combined by the geometric mean using AFNI's `3dFWHMx` tool. Since this value varies by voxel size, it is normalized by the geometric mean of the voxel dimensions to render it comparable across different acquisition parameters. Head motion and technical factors will blur the spatial details of an image and increase the spatial smoothness. Since image smoothness is bound by the voxel size, the minimum value of this value should be 1, and values closer to this minimum are better.

2.1.1.5 Percentage of artifact voxels (Qi1)

Since thermal noise should not exhibit spatial correlation, any spatial structure in the background is assumed to reflect artifacts such as ghosts and RF banding. The percentage of artifact voxels, or the Quality Index (QI1, Mortamet et al. (2009)), is a measure of the proportion of voxels in the background that contain artifacts to the number of voxels in the background (the air). Artifacts in background regions that are not proximal to the brain are not considered critical to image quality and are excluded from the calculation.

Artifact voxels are identified from the background using the following procedure Mortamet et al. (2009):

1. A background mask is calculated from the inverse of a head mask and further constrained to voxels that are superior to an oblique plane that connects the bottom of the forehead to the nape of the neck
2. The mode of background voxel intensities is used as a threshold to exclude low intensity noise values from consideration
3. A modified morphological opening operation that consists of an erosion using a 3D cross followed by a dilation is applied to the remaining voxels to remove unconnected voxels (i.e. those that are not adjacent to other supra-threshold voxels)
4. Remaining voxels are labeled as belonging to artifact

QI1 is the percentange of background voxels (V_B) that are classified as artifacts (V_A), (6).

$$QI1 = \frac{|V_A|}{|V_B|} \tag{6}$$

QI1 is calculated only on structural data, and the closer this number is to zero, the better.

2.1.1.6 Signal-to-Noise Ratio (SNR)

Signal-to-Noise Ratio is a ubiquitous measure of how well image features of interest are differentiable from obscuring variation. The definition of signal and noise can vary based on the intent for the images being evaluated, but for structural MRI it is defined as the mean of a homogeneous region of image over the variation in a region of the background. To simplify the automatic calculation of this measure, it is calculated in QAP as the ratio between the mean of the gray matter signal ($\overline{GM}$) and the standard deviation of the signal intensity of all background voxels (σ_b) (Magnotta and Friedman (2006)):

$$SNR = \frac{\overline{GM}}{\sigma_b} \tag{7}$$

SNR is calculated for both anatomical and functional data, and the greater this number is, the better.

2.1.1.7 Ghost-to-Signal Ratio (GSR)

Ghosts in MRI images can arise from a variety of technical sources, such as gradient calibration and sequence parameters, as well as patient motion. For fMRI data, the ghosts appear in the phase encoding direction making them easy to locate and measure. GSR is the difference between the mean voxel intensity of background regions where ghosts are likely to occur ($\overline{V_G}$) and the mean voxel intensity of the remainder of background voxels ($\overline{V_B}$), divided by the mean voxel intensity from within the brain ($\overline{V_F}$) ((Giannelli et al., 2010)):

$$GSR_j = \frac{\overline{V_G} - \overline{V_B}}{\overline{V_F}} \tag{8}$$

GSR is only calculated on functional data, and the closer this value is to zero this value is, the better.

2.1.2 Measures of temporal quality

add change to these metrics, inclusion of Mean/Std.Dev, etc.

2.1.2.1 Mean Standardized DVARS

Head motion and other scanner instabilities may induce large global intensity variations (spikes) in fMRI time series. DVARS measures these variations from the spatial standard deviation of frame-to-frame difference images (Power et al., 2012). DVARS in its original form is effected by the temporal autocorrelation in the data, which makes it impossible to compare its values meaningfully between different acquisition strategies and sites. Standardized DVARS accounts for this autocorrelation, resulting in a more absolute measure of DVARS that is comparable between datasets ((Nichols, 2013)):

$$DVARS = \frac{1}{P}\sum_{p}^{P}\frac{\sqrt{\frac{1}{N}\sum_n (V_{n,p} - V_{n,p-1})^2}}{\sqrt{\frac{1}{N}\sum_n 2(1-\rho_n)\sigma_n^2}} \tag{9}$$

$V_{n,p}$ being the intensity of the p^{th} observation of the n^{th} voxel, N being the total number of voxels and P being the number of observations (TRs), σ_n^2 is temporal variance, and ρ_n is the one-lag temporal auto-correlation. The closer this value is to zero, the better.

2.1.2.2 Outlier Detection

Similar to DVARS, the goal of outlier detection, defined by AFNI's `3dToutcount`, is to quantify the presence of spikes in the fMRI time series. This measure is defined as the fraction of outlier voxels per volume averaged across the fMRI time series. Voxels are determined to be outliers if their intensity exceeds a threshold defined by:

$$\sqrt{\frac{\pi}{2}}\Phi^{-1}\left(\frac{0.001}{P}\right) MAD \tag{10}$$

where Φ^{-1} is the inverse of the reversed Gaussian cumulative distribution function and P is the total number of observations (time points). The mean absolute deviation (MAD) of the voxel time series is calculated from:

$$MAD = med_p(|V_{n,p} - \widetilde{V_n}|) \tag{11}$$

where med_p is the median operator over observations, $\widetilde{V_n}$ is the median observation of voxel n, and $V_{n,p}$ is the p^{th} observation of voxel n. The closer this value is to zero, the better.

2.1.2.3 Median Quality Index (MQI)

The Median Quality Index, as defined by AFNI's `3dTqual` tool, is a multivariate analog of the previously described outlier detection. A quality index for each fMRI volume is calculated from one minus its spatial Spearman's rank correlation with the median volume:

$$QI_p = 1 - \frac{1}{N-1}\sum_{n=1}^{N}\frac{(R_{n,p} - \overline{R_p})(\widetilde{R_n} - \overline{\widetilde{R}})}{\sigma_{R_p}\sigma_{\widetilde{R}}}, \tag{12}$$

where $R_{n,p}$ is the rank of the n^{th} voxel in the p^{th} volume, $\overline{R_p}$ is the mean rank of all voxels in the p^{th} volume, $\widetilde{R_n}$ is the rank of the n^{th} voxel in the median volume, $\overline{\widetilde{R}}$ is the mean rank of voxels in the median volume, and σ_{R_p} and $\sigma_{\widetilde{R}}$ are the standard deviation of the ranks of voxels in the p^{th} and median volumes respectively. MQI ($\widetilde{QI}$) is the median quality index across the volumes of the fMRI dataset. This value varies between zero and two, and the closer to zero this value is, the better.

2.1.2.4 Quality Index (QI) Percent Outliers

This is the percent of total volumes whose quality index (described in 2.1.2.3) are statistical outliers as determined by the criterion:

$$QI_p \geq \widetilde{QI} \pm MAD(QI) \tag{13}$$

where $\widetilde{QI}$ is the median quality index (described in 2.1.2.3) and $MAD(QI)$ is the median absolute deviation of the quality indices of all volumes in the dataset and is calculated similar to Eqn. 11. The fewer the number of outliers (the smaller the percentage), the better.

2.1.2.5 Global Correlation (GCOR)

Certain types of nuisance variation in fMRI data will increase the correlation between voxel time series. This is particularly true for head motion and physiological signals (respiration and heart beat) and can be exaggerated by their interaction with other imaging parameters (e.g. parallel imaging, slice gap, multi-band imaging) (Saad et al., 2013). GCOR (γ) quantifies this effect by the average correlation between every pair of in-brain voxels in the fMRI data:

$$\gamma = \frac{1}{N^2} \sum_{m=1}^{N} \sum_{n=1}^{N} \rho_{m,n} \tag{14}$$

where N is the number of voxels and $\rho_{m,n}$ is the Pearson's correlation between the P length time series for the n^{th} and m^{th} voxels V_n and V_m respectively, each having mean $\overline{V_n}$ ($\overline{V_m}$) and standard deviation σ_n (σ_m):

$$\frac{1}{P-1} \sum_{p=1}^{P} \frac{(V_{n,p} - \overline{V_n})(V_{m,p} - \overline{V_m})}{\sigma_n \sigma_m}. \tag{15}$$

Lower values of GCOR are preferred.

2.1.2.6 Mean Root Mean Square Deviation (MeanRMSD)

Root mean square deviation (RMSD) is a measure of frame-to-frame motion in an fMRI time series that summarizes the three translations (x, y, z) roll (α), pitch (β), and yaw (γ) estimated from the linear co-registration between each volume of the fMRI time series and a reference volume (Jenkinson, 1999). A 4×4 transform matrix T_t that maps volume t to the reference volume can be calculated from these six parameters using:

$$T_t = \begin{bmatrix} 1 & 0 & 0 & x \\ 0 & 1 & 0 & y \\ 0 & 0 & 1 & z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha & 0 \\ 0 & -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\gamma & \sin\gamma & 0 & 0 \\ -\sin\gamma & \cos\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{16}$$

Using these transforms, the distance $M_{t,t+1}$ between volume t and volume $t+1$ can be decomposed into a 3×3 matrix A and 3×1 vector b:

$$T_{t+1} T_t^{-1} - I = \begin{bmatrix} A & b \\ 0 \; 0 \; 0 & 0 \end{bmatrix}. \tag{17}$$

With these and the center of the volume c, the root mean square deviation between the two frames is calculated from (using the notation in Yan et al., 2013):

$$RMSD = \sqrt{\frac{1}{5} R^2 Tr[A^T A] + (b + Ac)^T (b + Ac)}, \tag{18}$$

where R is the radius of the head, which is assumed to be 80mm. MeanRMSD is the mean RMSD calculated for each consecutive pair of volumes in an fMRI series. We chose MeanRMSD over other measures of frame-to-frame motion (e.g. Power et al., 2012; Dijk et al., 2012) due to our previous observations that it is a more accurate estimate of the motion seen at the voxel level (Yan et al., 2013). The closer to zero this measure is, the better.

2.2 The QAP Python Toolbox

The processing procedures required by QAP, such as image segmentation, registration, and mask generation are all accomplished using components from AFNI (Cox, 1996). These tools are pipelined together with QAP-specific python modules using Nipype (Gorgolewski et al., 2016b) to simplify the processing of very large datasets using high performance computing system such as multicore workstations and clusters. Other benefits of Nipype include provenance tracking and a mechanism that enables a pipeline to be restarted after a change in configuration or error and only recompute the affected pipeline steps. The amount of processing required for calculating the metrics has been minimized so that they focus on data quality rather than the quality of the algorithms used to perform the processing. But since some processing (e.g., segmentation, alignment, and masking) is unavoidable we recommend that the same algorithms be used on all data for which QA measures will be compared.

2.2.1 System requirements and installation

QAP can run on any system that supports Python along with the AFNI neuroimaging toolset. This currently includes most *nix platforms and Mac OS X. QAP can run on fairly modest workstations and also supports parallel execution on multicore workstations and Sun Grid Engine (Gentzsch, 2001), Condor (Thain et al., 2005), PBS (Jones, 2002), and Slurm (Jette et al., 2002) based clusters. QAP can be installed using standard python package installation tools (e.g. pip) and is available preinstalled on a free-to-use Amazon Machine Image. Extensive documentation on installing and using QAP are available at its webpage[1]. If it is ready for that, we could mention the docker image here

here a sentence to connect to the next section: why we need a pipeline to produce the intermediate results on which these measures are computed

Additionally, a “Resource of QAP measures” calculated for the ABIDE and CoRR datasets is available on the Preprocessed Connectomes Project’s main page (link here?) as a spreadsheet matching the participant IDs with their scans’ appropriate quality measures. OE: isnt this last paragraph a result itself?

2.2.2 Pipeline execution

In Nipype terminology, the QAP pipelines are composed of nodes that represent different AFNI tools or Python functions that implement the required image processing steps and calculate the quality measures. Nodes are connected together by their respective input and output files, most of which are intermediary files that can be deleted after the calculation has completed without affecting the outputs. Nodes are executed in a working directory, which contains log files that provide more details on the processing that occurred and any intermediary files that are produced generated. Once execution has completed the final results are copied to an output directory, and the working directories can be deleted. If QAP is restarted and the working directory already exists (e.g. has not been deleted), it will use any already-computed intermediaries that are in the directory, rather than recomputing them. This enables a warm-restart capability to allow the user to recover from an error or reconfiguration without having to recompute all of the processing steps.

QAP has been designed to avoid unnecessary dependencies between different datatypes; estimating QAP on functional data does not require any information from an anatomical scan acquired during the same session and likewise temporal and spatial quality assessment on the functional data are kept separate. Although this may limit some of the processing that can be performed, it was done to avoid contaminating the estimated quality of functional data with poor quality anatomical data. Applying this criterion allowed

[1] http://preprocessed-connectomes-project.github.io/quality-assessment-protocol/#installing-the-qap-package

the QAP to be broken into three different pipelines: `qap_anatomical_spatial.py` for calculating spatial measures on anatomical scans, `qap_functional_spatial.py` for calculating spatial measures on functional (EPI) scans, and `qap_functional_temporal.py` for calculating temporal measures on functional scans.

Each pipeline script requires a dataset list and a configuration file in the easy-to-construct Python YAML format. The same dataset list can be used for all three scripts and contains file paths to the input data to be assessed. At the very least, this list must contain paths to the raw anatomical or functional data, but may also contain intermediary files that will be used to bypass their corresponding pipeline nodes. In this way, the user can override any of the processing steps in the pipeline with precomputed data from their preferred implementations. QAP includes scripts which can auto-generate dataset lists from a directory or data arranged in the BIDS format (Gorgolewski et al., 2016a). Currently QAP only supports data in compressed or uncompressed NIfTI files (Cox et al., 2004). The configuration file contains a small collection of settings for the pipeline, such as the location of the working and output directories, configuring parallel execution, and the path to needed template files.

Once a script is executed, a pipeline builder is invoked to dynamically construct a Nipype pipeline using information in the dataset list and configuration files. A resource pool is populated with all of the files specified in the data list along with any files that are found in the working directory. The pipeline is then constructed from the bottom up by first looking for a needed intermediary in the resource pool, if it is not found then the pipeline node that generates the file is inserted, and then its inputs are searched for, and so on. The resulting pipeline is submitted to a Nipype execution scheduler that tracks the data dependencies between pipeline steps to identify those that can be run in parallel. By combining the processing of multiple datasets into a single pipeline, QAP takes advantage of parallelizability available both between processing steps performed on different datasets and within steps performed on the same dataset to maximize throughput. The scheduler takes into account estimates of the amount of memory and processors in use and estimates of the resources required by ready-to-run pipeline steps when making scheduling decisions to avoid overloading the system's resources.

When using a cluster, QAP separates the data list into jobs that are submitted to the scheduler such that a separate instance of the QAP script is run on each cluster node. This ensures that all of the processing for a dataset occurs on the same node, minimizing the data transfer between nodes to improve computation speed. When running on a cluster we recommend to use data storage that is local to each cluster node, rather than a network share, for the working directory. This will reduces delays due to network transfers and competition between cluster nodes for access to the same resource. A disadvantage of this is that it makes it harder to use the ”warm restart” functionality of QAP, but we have found that the pipelines are quick enough that this is not that high of a penalty compared to the aforementioned network costs. Once a pipeline is completed the output statistics can be copied back to shared storage.

The output of the QAP scripts is a CSV file per dataset that includes the calculated measures for that dataset. A separate script is provided to combine CSVs across datasets.

2.2.3 Pipeline steps - Anatomical

The anatomical spatial pipeline uses AFNI's `3drefit` to deoblique the raw anatomical images, and AFNI's `3dresample` to reorient the deobliqued image to RPI orientation. This resulting image is used throughout the QAP spatial anatomical metric calculations as the source of raw anatomical data.

The pipeline uses 3dSkullStrip and 3dCalc to remove the skull from the deobliqued, reoriented anatomical scan, providing a brain-only image to be used in tissue segmentation, performed using 3dSeg. This generates gray matter, white matter, and CSF masks for use in the calculation of CNR and Cortical Contrast.

A whole-head mask is created from the deobliqued, reoriented anatomical scan using 3dCalc, with an appropriate threshold selected by 3dClipLevel. A sequence of six dilations and erosions are performed using 3dmask_tool in order to fill in gaps within the mask.

Because some of the spatial metrics rely on the signal intensity in the background of the image, motion artifacts present in the area near and under the participant's mouth and nose can introduce error. In the process of generating an appropriate foreground and background mask of the anatomical data, AFNI's 3dAllineate is used to perform linear anatomical registration to a user-specified template image.

The resulting affine matrix is used to calculate and draw a mask segment covering the area of background in the anatomical image directly in front of the participant's mouth and chin. 3dcalc is used to combine the whole-head mask with this slice. An inversion of this mask results in the background mask, which is used in the FBER, Qi1, SNR and CNR spatial measures. A ”skull-only” mask is also generated by subtracting the slice mask from the whole-head mask, which is used in the FBER calculation.

2.2.4 Pipeline steps - Functional

The functional pipelines start with using AFNI's 3drefit and 3dresample to deoblique and reorient the functional images. Motion correction is then run using 3dvolreg to obtain the coordinate transformation, which is used to calculate the Root Mean Square Deviation metric for the functional temporal pipeline.

A functional brain mask is generated using 3dAutomask. This mask is used in conjunction with the deobliqued, reoriented functional timeseries to calculate the other functional temporal metrics Outliers, Quality, and Global Correlation (GCOR). This mask is then inverted using 3dcalc and used to calculate Out-of-Brain (OOB) Outliers, a measure of intensity spikes residing outside of the brain. Outliers and OOB Outliers are calculated using AFNI's 3dToutcount, and Quality is calculated using 3dTqual. The mean and standard deviation of each temporal metric is also calculated and reported, along with the median and inter-quartile range (IQR).

The timeseries is also averaged into the one-volume mean functional image for spatial metrics using AFNI's 3dTstat. This mean functional image is used with the functional brain mask to calculate EFC, FWHM, and Ghost-to-Signal Ratio. The background mask, which is an inversion of the functional brain mask, is used with the mean functional image and the original brain mask to calculate FBER.

2.2.5 Graphical Reports

Since no automated quality assessment is perfect, QAP generates a series of graphical reports to optimize the process of eyeballing the images and quickly detect outliers across the different metrics. Two categories of graphical reports are produced: individual reports (one per subject in the data pool) and group reports (showing the distribution of subjects across measures). The individual reports are structured as follows: 1) a section documenting the measures included in the corresponding report, along with an identifier of the subject and description of the number of sessions and runs included; 2) mosaic views of axial slices of images of interest; and 3) violin plots showing the distributions of measures for all subjects, indicating the location of the corresponding subject in each distribution. For the anatomical spatial protocol, the image of interest for the mosaic view is the original T1-weighted image. For the functional reports, the temporal image of interest is the the tSNR map (average of the SNR map of each timepoint) and the spatial image of interest is the averaged signal along time.

OE: comments on this section
1. I would split out a whole data section, including this description of data would be better in a table, where the numbers of subjects per modality (func, anat) are easily identified, along with proper citation (papers or links)
2. I would remove the experiments from here (what we evaluated) and place in a section where it is more clear that this is what we are looking at after extracting the metrics.

??? says three expert raters here, but says four raters down below in Statistical Analysis !!

The Preprocessed Connectomes Project Quality Assessment Protocol was used to calculate spatial and temporal quality measures on the (now 1,101 structural and 1,163 functional what does this mean??) 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 four 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 intra-class correlation.

2.3 Methods for assessing the quality of ABIDE data

2.4 Statistical Analysis

2.4.1 Correlations between measures

Pearsons correlation coefficients were computed to assess the relationship between measures for each dataset separately, ABIDE and CoRR, and were summarized into table 1 and figure 2. Note that we considered the first scan of the first session for each participant to compute correlations in CoRR since CoRR consists of multiple scans of multiple sessions for each subject.

2.4.2 Test-retest of the measures (for CoRR)

We were able to evaluate the test-retest reliability for each measure using CoRR dataset since CoRR has multiple scans of multiple sessions for each subject. The intra-class correlations (ICC: Shrout and Fleiss (1979)) have been computed for each site by

$$ICC = \frac{MS_b - MS_w}{MS_b + MS_w} \tag{19}$$

where MS_b is the between-subject mean square and MS_w is the within-subject mean square for each measure.

2.4.3 Relationship of measures with hand assessments

Manually applied structural data quality sources are available for ABIDE. Hence, we evaluated the relationship of measures with hand assessments. Four individual reviewers scored each image. When three or more reviewers agreed that the quality of the image is okay, we considered the image as a “success” (or pass). Then, we applied the logistic regression to hand assessments against quality measures.

2.4.4 Relationship of measures with scanning parameters

The associations between quality measures and scanning parameters, with respect to all measures of interest, were based on mixed effects models analyses (Diggle et al., 2013). Scanning parameters include type of scanner, slice gap, slice acquisition, flip angle, TE, TR, slice thickness, voxel area, duration of scan N (timepoints). Scanner, slice gap and slice acquisition were considered factors with multi-level. Random site effects were included in the models since observations are nested in sites.

3 RESULTS

3.1 Summary

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.

3.2 Correlations between measures

We notice that anatomical measures more correlated each other than functional measures, and correlations between measures in ABIDE are similar to correlations in CoRR.

Table 1. Correlations between anatomical measures. CoRR in the upper triangular, ABIDE in the lower triangular. *indicates p-value less than 0.05

Anatomical	CNR	EFC	FBER	FWHM	Qi1	SNR
CNR		-0.506*	0.316*	-0.111	-0.489*	0.733*
EFC	-0.458*		-0.126	-0.151	0.572*	-0.562*
FBER	0.436*	-0.286*		-0.038	-0.197	0.444*
FWHM	-0.056	0.109	0.056		-0.082	-0.137
Qi1	-0.392*	0.372*	-0.238	0.073		-0.549*
SNR	0.760*	-0.61*	0.675*	-0.112	-0.448*	

Functional	EFC	FBER	FWHM	SNR	DVARS	Mean RMSD	Quality	GCOR
EFC		-0.264*	0.140	-0.707*	-0.036	-0.007	-0.820	-0.056
FBER	-0.134		0.053	0.838*	0.100	0.070	0.290*	0.017
FWHM	-0.023	0.168		-0.036	-0.175	0.013	-0.030	-0.021
SNR	-0.586*	0.845*	0.118		-0.047	0.070	0.624*	0.043
DVARS	0.012	0.015	-0.108	-0.016		-0.020	-0.116	0.197
Mean FD	-0.029	0.129	0.143	0.139	-0.140		0.162	-0.014
Quality	-0.693*	0.234	0.106	0.531	-0.153	0.310*		-0.042
GCOR	-0.098	0.066	0.035	0.092	0.207	0.060	0.046	

Table 2. Logistic regression results — Anatomical

Measure	Estimate	Std Err	p-value
CNR	0.370	0.058	0.000
EFC	2.463	2.369	0.298
FBER	0.002	0.001	0.009
FWHM	0.083	0.126	0.513
Qi1	-7.484	1.064	0.000
SNR	-0.179	0.039	0.000

3.3 Test-retest of the measures (for CoRR)

Figure 3 shows the boxplots of ICCs of each site for each measure. Note that variances of ICCs of functional measures are less than of anatomical measures. EFC, FBER, SNR, Percent FD for functional measures have very high ICCs average over 0.75 while ICC of EFC for anatomical is average over 0.75

3.4 Relationship of measures with hand assessments

Table 2 summarizes the results. Figure 4 shows boxplots of most discriminative measures vs. hand assessments. CNR, QI1, SNR are significant predictors of hand assessment in anatomical while all measures except FBER, SNR are significant in functional.

3.5 Relationship of measures with scanning parameters

4 CONCLUSION

This is where the conclusion will go.

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.

Table 3. Logistic regression results — Functional

Measure	Estimate	Std Err	p-value
EFC	7.136	2.823	0.011
FBER	-0.041	0.020	0.043
FWHM	1.864	0.326	0.000
SNR	0.369	0.145	0.011
Quality	4.846	3.434	0.158
RMSD	-2.139	0.480	0.000
DVARS	1.809	0.754	0.016
GCOR	3.789	1.016	0.000

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 Text.

Funding: Text Text Text Text Text Text Text Text.

Table 4. Anatomical spatial - Regression analysis for the relationship of measures with scanning parameters. Estimated coefficients of each parameter are reported by measures. * indicates p-value less than 0.05.

Parameter / measure		CNR	EFC	FBER	FWHM	Qi1	SNR
Intercept		-153.400	1.568*	-8160.880	-11.093	0.956	-116.510
Scanner	GE MR750	11.463	-0.082	1977.000	-8.292	-0.030	-20.262
	GE Sig	-4.534	-0.124	643.650	1.379	-0.188	21.553
	Phillips Achieva	45.799*	-0.383*	2386.040	9.017	-0.174	55.190
	Phillips Intera	15.405*	-0.146*	410.430	-0.819	-0.168*	11.042
	Siemens Allegra	15.587	-0.044	2564.100	-7.555	-0.270*	-15.226
	Siemens Tim TRIO	4.096	-0.020	1110.120	-4.282	-0.213*	-15.063
	Siemens Verio	0.000	0.000	0.000	0.000	0.000	0.000
Slice Gap	0	5.776	-0.037	232.290	2.314	0.063	11.518
	1	0.000	0.000	0.000	0.000	0.000	0.000
Slice Acquisition	int+	0.000	0.000	0.000	0.000	0.000	0.000
	seq+	-2.640	0.215	1848.390	-11.416	-0.060	-39.822
	seq-	0.000	0.000	0.000	0.000	0.000	0.000
Flip Angle		-0.574*	0.000	-99.215	0.176*	0.003	-0.168
TE		2.450*	-0.006	231.930	-0.253	-0.017*	1.135
TR		0.022*	0.000*	2.945	-0.003	0.000*	0.008
Slice Thickness		5.918	-0.083*	-640.360	2.313	-0.038	7.966
Voxel Area		1.457	-0.005	54.819	0.676	0.019*	3.021
N Timepoints		0.260*	-0.002*	21.238	0.020	-0.001*	0.200

REFERENCES

Atkinson, D., Hill, D. L., Stoyle, P. N., Summers, P. E., and Keevil, S. F. (1997). Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Trans Med Imaging* 16, 903–910

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173

Cox, R. W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C. J., et al. (2004). A (sort of) new image data format standard: NIfTI-1. In *Proceedings of the 10th Annual Meeting of Organisation of Human Brain Mapping (2004)* (Budapest, Hungary)

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667

Dijk, K. R. V., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity {MRI}. *NeuroImage* 59, 431 – 438. doi:http://dx.doi.org/10.1016/j.neuroimage.2011.07.044. Neuroergonomics: The human brain in action and at work

Table 5. Functional spatial - Regression analysis for the relationship of measures with scanning parameters. Estimated coefficients of each parameter are reported by measures. * indicates p-value less than 0.05.

Parameter / measure		EFC	FBER	FWHM	SNR
Intercept		1.980	-589.070	-3.098	-27.158
Scanner	GE MR750	-0.093*	103.830	0.367	3.422
	GE Sig	-0.261*	46.691	1.474*	3.133
	Phillips Achieva	-0.591*	230.290*	2.720*	11.073*
	Phillips Intera	-0.226*	93.450*	0.733*	3.676*
	Siemens Allegra	-0.046	94.377	-0.515*	3.757
	Siemens Tim TRIO	-0.035	44.902	-0.271	2.026
	Siemens Verio	0.000	0.000	0.000	0.000
Slice Gap	0	-0.014	-8.429	0.597*	-0.354
	1	0.000	0.000	0.000	0.000
Slice Acquisition	int+	0.000	0.000	0.000	0.000
	seq+	0.329*	-57.272	-2.077*	-4.420
	seq-	0.000	0.000	0.000	0.000
Flip Angle		0.000	-1.340	-0.001	-0.030
TE		-0.016*	10.653*	0.007	0.582*
TR		0.000	0.062*	0.000*	0.002*
Slice Thickness		-0.108*	28.465	0.465*	1.917
Voxel Area		-0.027*	9.621	0.126*	0.493
N Timepoints		-0.001*	0.530	0.006*	0.025

Friedman, L. and Glover, G. H. (2006). Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *Neuroimage* 33, 471–481

Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., et al. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Hum Brain Mapp* 29, 958–972

Gentzsch, W. (2001). Sun grid engine: Towards creating a compute power grid. In *Proceedings of the 1st International Symposium on Cluster Computing and the Grid* (Washington, DC, USA: IEEE Computer Society), CCGRID '01, 35–

Giannelli, M., Diciotti, S., Tessa, C., and Mascalchi, M. (2010). Characterization of Nyquist ghost in EPI-fMRI acquisition sequences implemented on two clinical 1.5 T MR scanner systems: effect of readout bandwidth and echo spacing. *J Appl Clin Med Phys* 11, 3237

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* 80, 105–124

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016a). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* 3, 160044. doi:10.1038/sdata.2016.44

Table 6. Functional temporal - Regression analysis for the relationship of measures with scanning parameters. Estimated coefficients of each parameter are reported by measures. * indicates p-value less than 0.05.

Parameter / measure		DVARS	Quality	Mean RMSD	Perc FD	Num FD	GCOR
Intercept		2.838	0.060	2.274	36.128	57.702	-8.796
Scanner	GE MR750	-0.348	0.019	0.233	8.094	11.569	-0.423*
	GE Sig	-0.309	-0.021	-0.458	-7.818	-11.675	1.581*
	Phillips Achieva	-0.311	-0.043	-0.804	-16.029	-26.390	3.002*
	Phillips Intera	-0.058	-0.015*	-0.206	-3.967	-8.222	0.723*
	Siemens Allegra	-0.349	0.026*	0.239	7.412	10.758	-0.361
	Siemens Tim TRIO	-0.266	0.015	0.146	5.404	7.291	-0.210
	Siemens Verio	0.000	0.000	0.000	0.000	0.000	0.000
Slice Gap	0	0.105	-0.014*	-0.119	-3.159	-5.460	0.222*
	1	0.000	0.000	0.000	0.000	0.000	0.000
Slice Acquisition	int+	0.000	0.000	0.000	0.000	0.000	0.000
	seq+	0.118	0.046	0.771	17.686	27.689	-2.604*
	seq-	0.000	0.000	0.000	0.000	0.000	0.000
Flip Angle		0.007	0.000	-0.007	-0.159	-0.266	0.006
TE		-0.048	0.001	-0.013	-0.126	-0.141	0.071*
TR		0.000	0.000	0.000	0.003	0.005	0.000*
Slice Thickness		-0.098	-0.008	-0.233	-4.183	-6.971	0.936*
Voxel Area		0.016	-0.003*	-0.039	-0.858	-1.276	0.111*
N Timepoints		-0.002	0.000	-0.001	-0.023	-0.025	0.009*

Gorgolewski, K. J., Esteban, O., Burns, C., Ziegler, E., Pinsard, B., Madison, C., et al. (2016b). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. doi:10.5281/zenodo.50186

Jenkinson, M. (1999). *Measuring transformation error by RMS deviation.* Internal technical report TR99MJ1, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain

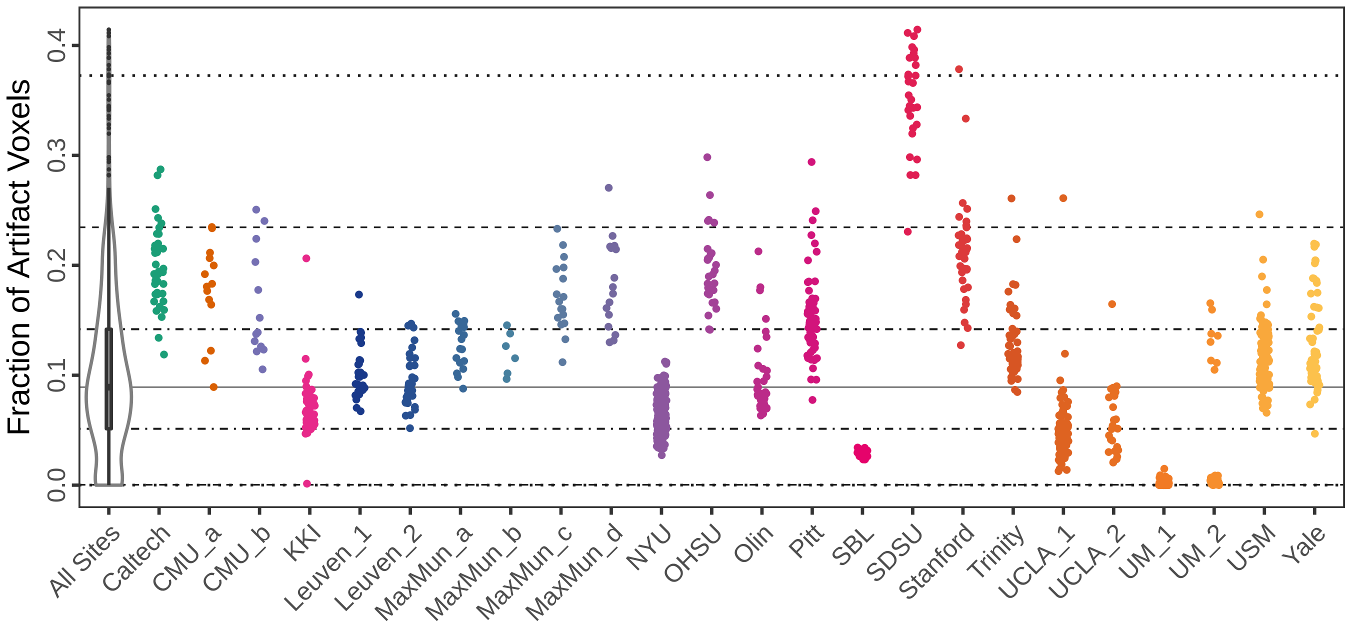
Jette, M. A., Yoo, A. B., and Grondona, M. (2002). Slurm: Simple linux utility for resource management. In *In Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP) 2003* (Springer-Verlag), 44–60

Jones, J. P. (2002). Beowulf cluster computing with windows (Cambridge, MA, USA: MIT Press), chap. PBS: Portable Batch System. 363–383

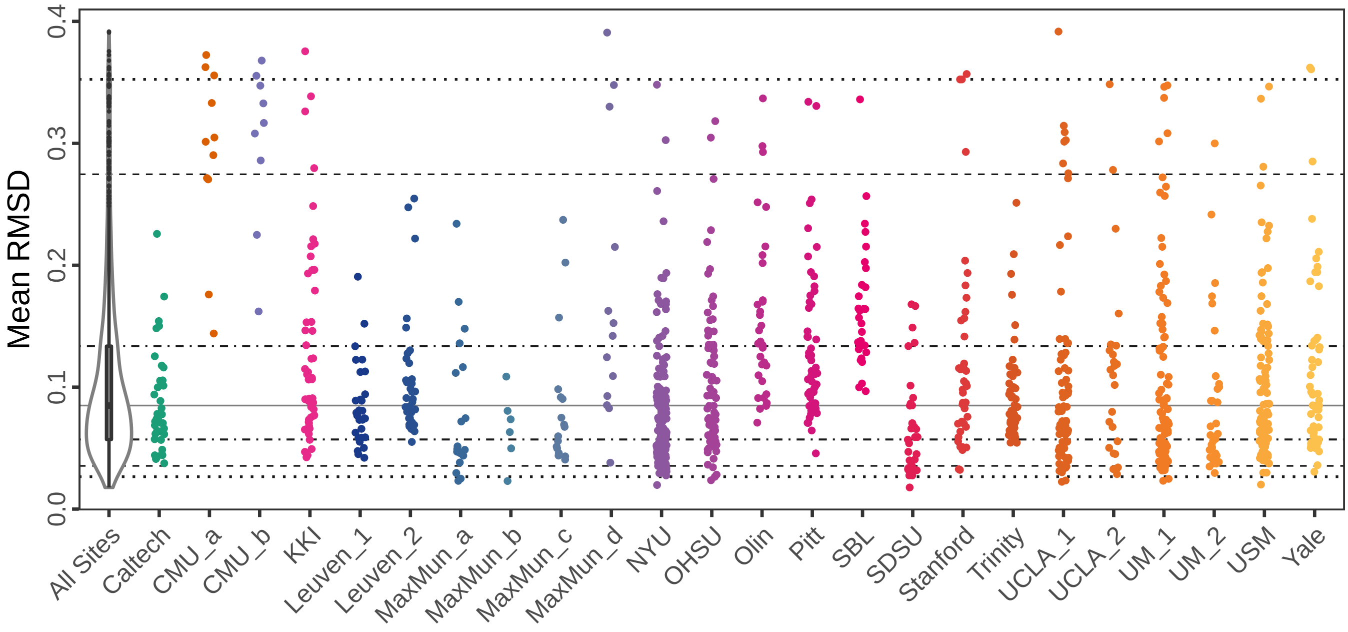
Magnotta, V. A. and Friedman, L. (2006). Measurement of Signal-to-Noise and Contrast-to-Noise in the fBIRN Multicenter Imaging Study. *J Digit Imaging* 19, 140–147

Mortamet, B., Bernstein, M. A., Jack, C. R., Gunter, J. L., Ward, C., Britson, P. J., et al. (2009). Automatic quality assessment in structural brain magnetic resonance imaging. *Magn Reson Med* 62, 365–372

Nichols, T. (2013). Notes on creating a standardized version of DVARS. [http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/



(a) Structural MRI fraction of artifact voxels.

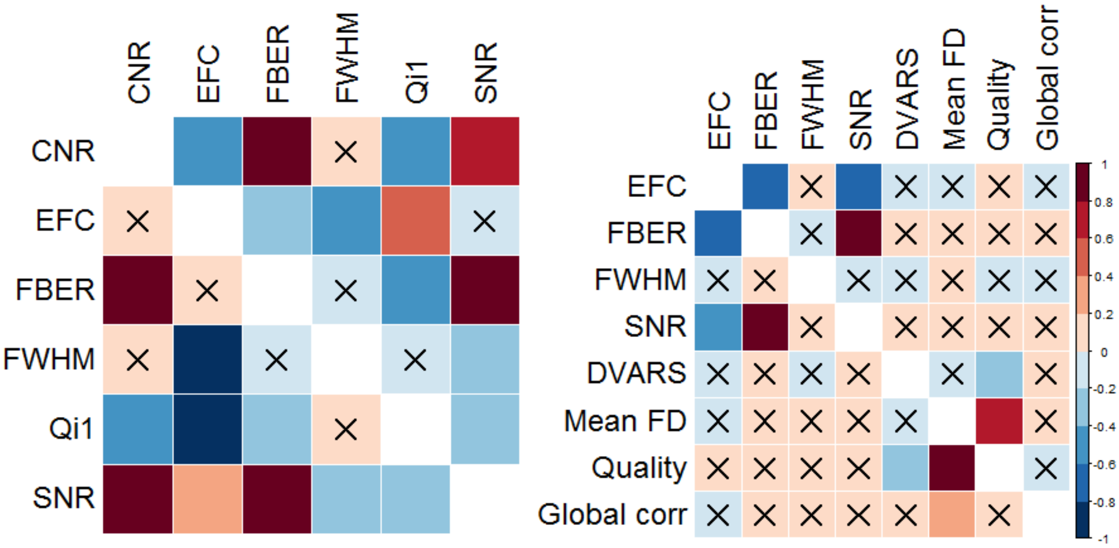


(b) Functional MRI mean root-mean-square deviation.

Figure 1: Examples of distributions for QAP measures calculated on ABIDE.

nichols/scripts/fsl/standardizeddvars.pdf; last viewed Dec 4, 2015]

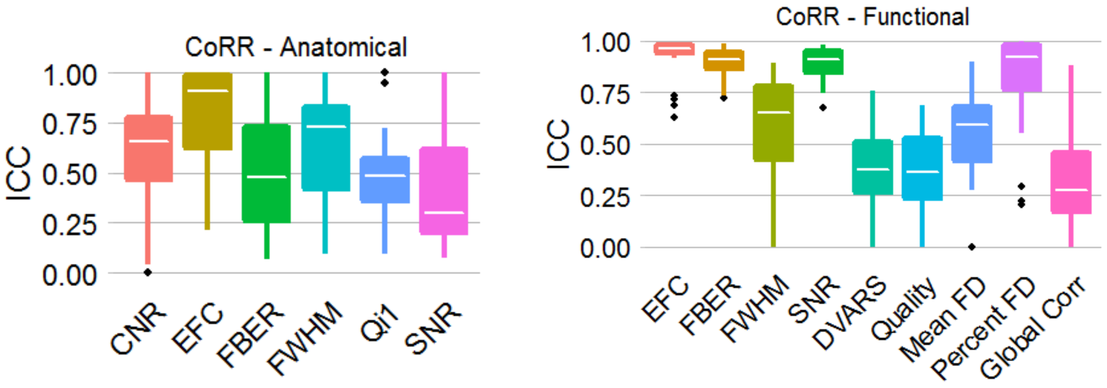
Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154



(a) Anatomical measures

(b) Functional measures

Figure 2: Correlogram of measures: ABIDE in the lower triangular, CoRR in the upper triangular, X=non-significant



(a) Anatomical measures

(b) Functional measures

Figure 3: Test re-test of measures for CoRR: Boxplots of ICCs of sites by quality measures.

Saad, Z. S., Reynolds, R. C., Jo, H. J., Gotts, S. J., Chen, G., Martin, A., et al. (2013). Correcting brain-wide correlation differences in resting-state FMRI. *Brain Connect* 3, 339–352

Shrout, P. and Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86, 420–428

Thain, D., Tannenbaum, T., and Livny, M. (2005). Distributed computing in practice: the condor experience. *Concurrency - Practice and Experience* 17, 323–356

Van Essen, D. C. and Ugurbil, K. (2012). The future of the human connectome. *Neuroimage* 62, 1299–1310

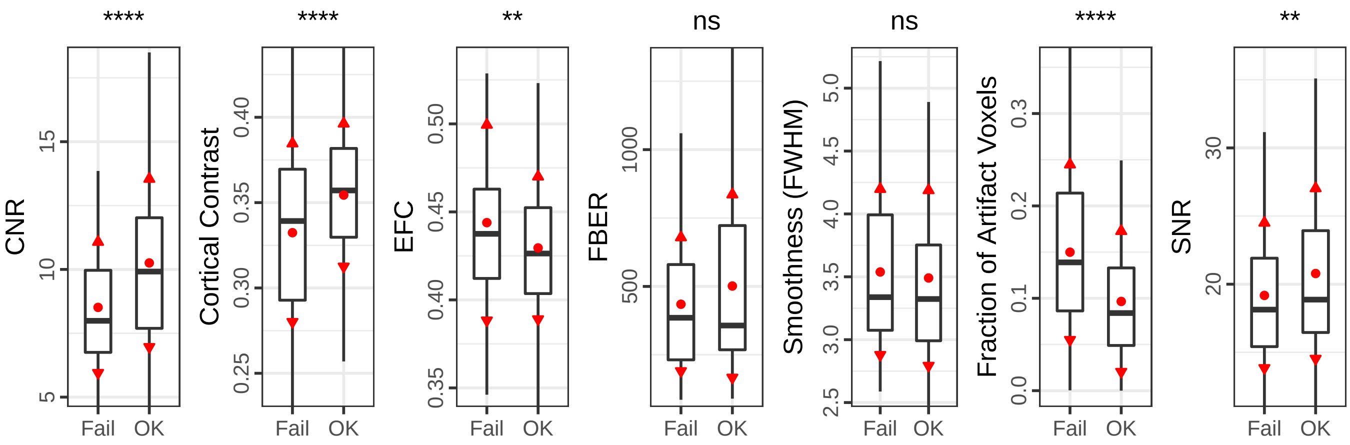


Figure 4: Structural MRI quality measures compared to results of visual inspection.

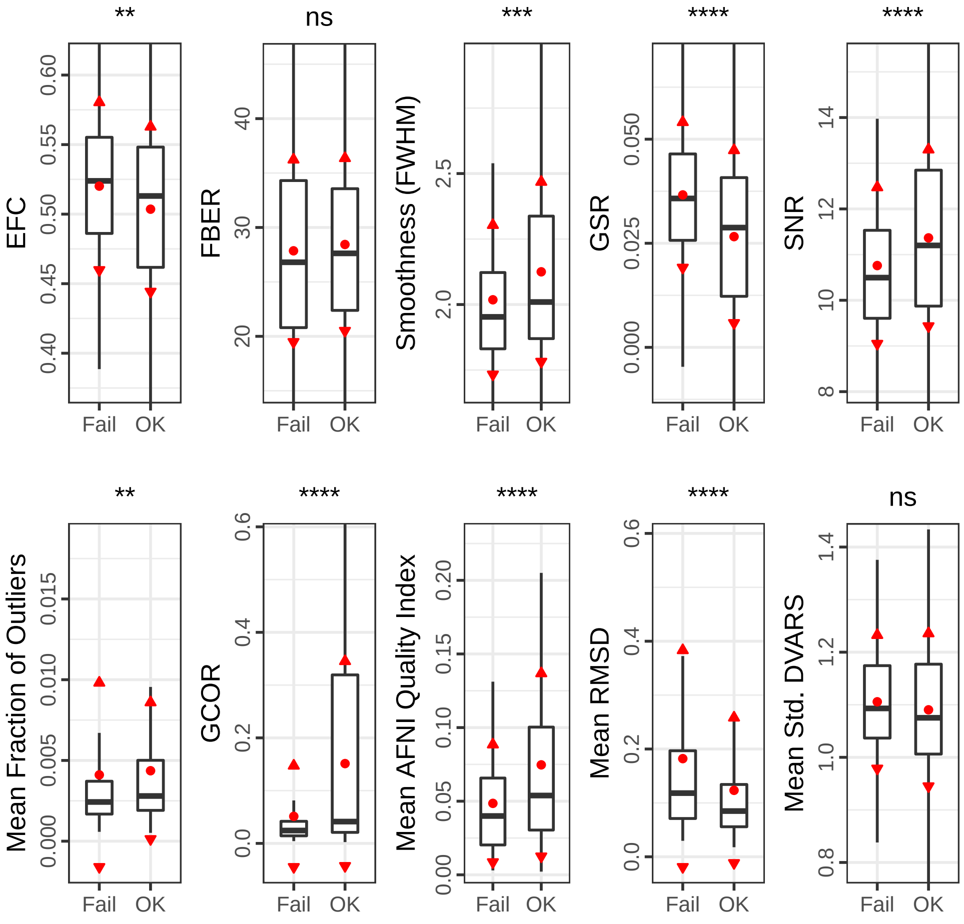


Figure 5: Functional MRI quality measures compared to results of visual inspection.

454 Yan, C. G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., et al. (2013). A
455 comprehensive assessment of regional variation in the impact of head micromovements on functional

connectomics. *Neuroimage* 76, 183–201

Zuo, X. N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci Data* 1, 140049

FIGURES

Figure 6. Enter the caption for your figure here. Repeat as necessary for each of your figures