Number of Words (body text): 8,131
Number of Figures: 10
Number of Tables: 4

Data-driven Assessment of Structural Image Quality

Adon F. G. Rosen,[a]* David R. Roalf,[a]* Kosha Ruparel,[a]
Jason Blake,[a] Kevin Seelaus,[a] Prayosha Villa,[a]
Phillip A. Cook,[b] Christos Davatzikos,[b,c] Mark A. Elliott,[b]
Angel Garcia de La Garza,[a] Efstathios D. Gennatas,[a] Megan Quarmley,[a]
J. Eric Schmitt,[a,b] Russell T. Shinohara,[d] M. Dylan Tisdall,[b]
R. Cameron Craddock,[e,f] Raquel E. Gur,[a,b] Ruben C. Gur,[a,b]
& Theodore D. Satterthwaite[a]

**Denotes shared first authorship.*

[a] Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA, USA
[b] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA, USA
[c] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia PA, USA
[d] Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia PA, USA
[e] Center for the Developing Brain, Child Mind Institute, New York, NY, USA
[f] Computational Neuroimaging Lab, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Old Orangeburg, NY, USA

Key words: motion, MRI, structural imaging, t1, development, artifact

Please address correspondence to:
Theodore D. Satterthwaite, M.D.
10th Floor, Gates Building
Hospital of the University of Pennsylvania
34th and Spruce Street
Philadelphia, PA 19104
sattertt@upenn.edu

ABSTRACT

Data quality is increasingly recognized as one of the most important confounders in brain imaging research. It is particularly important for studies of brain development, where age is systematically related to in-scanner motion and data quality. Prior work has demonstrated that in-scanner head motion biases estimates of structural neuroimaging measures. Yet, objective measures of data quality are not available for most structural brain images. Here we sought to identify reliable, quantitative measures of data quality for T1-weighted volumes, describe how such measures of quality relate to common measures of brain structure, and delineate how this in turn may bias inference regarding brain development in youth. Three highly-trained raters provided manual ratings for 1601 T1-weighted volumes acquired as part of the Philadelphia Neurodevelopmental Cohort. Expert manual ratings were compared to automated quality measures, which derived measures from the Preprocessed Connectomes Project's Quality Assurance Protocol (QAP). Generalized linear mixed-effects models using the automated quality measures were constructed in a training sample (n = 1067) to: 1) identify unusable images with significant artifacts, and 2) quantify subtle artifacts in usable images. These models were then tested in an independent validation dataset (n = 534). Results reveal that unusable images can be detected with a high degree of accuracy: a model including background kurtosis and skewness achieved an AUC of 0.95 in the training dataset and 0.94 in the independent validation dataset. While identification of subtle artifact was more challenging, an 8-parameter model achieved an AUC of 0.80 in the training dataset, and 0.92 in the validation dataset. Notably, quantitative measures of image quality were related to cortical thickness and gray matter density; measures of cortical volume were less affected by artifact. Furthermore, these quantitative measures of image quality demonstrated comparable or superior performance to estimates of motion derived from other imaging sequences acquired during the same protocol. Finally, data quality significantly altered structural brain maturation occurring during adolescent development. Taken together, these results indicate that reliable measures of data quality can be automatically derived from T1-weighted volumes, and that failing to control for data quality can systematically bias the results of studies of brain development.

INTRODUCTION

In-scanner motion and other imaging artifacts are increasingly appreciated as a source of bias in neuroimaging research. To the degree that in-scanner motion impacts image quality and is also related to characteristics of interest, such as age, clinical group, or treatment condition, it has the potential to systematically confound inference. While motion has long been a well-described methodological obstacle in medical imaging (Bellon et al., 1986; Smith and Nayak, 2010), and a known confound for task-related fMRI (Friston et al., 1996), there has been an acceleration of dedicated investigations following reports that even small amounts of in-scanner motion can bias studies of functional connectivity (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). Such work has re-invigorated interest regarding the potential impact of data quality on other imaging modalities, including T1-weighted neuroimaging of brain structure (Alexander-Bloch et al., 2016; Pardoe et al., 2016; Reuter et al., 2015; Savalia et al., 2017) and diffusion weighted imaging (Roalf et al., 2016; Yendiki et al., 2014).

Following initial work to assess the motion's impact (Atkinson et al., 1997), much subsequent work regarding structural imaging has focused on data quality issues driven by scanner and platform-related variation (Chen et al., 2014; Magnotta and Friedman, 2006; Styner et al., 2002; Woodard and Carley-Spencer, 2006). However, two published studies have used unique attributes of T1-weighted images to identify useable and unusable volumes. Mortamet et al. (2009) introduced a quality index (Qi) that accurately identified unusable volumes (AUC=0.93) collected as part of the Alzheimer's Disease Neuroimaging Initiative. Pizarro et al. (2016) developed statistics based on specific artifacts such as eye motion, ringing and tissue contrast. Combined in a multivariate approach these statistics classified unusable volumes with a classification accuracy of 80%. However, neither of these studies examined how these quality indices are related to measures of brain structure, or how quantitative indices of data quality might be used to account for biases in group level analyses of brain structure. This is particularly relevant given that measures of brain structure (including regional volumes, cortical thickness, and gray matter density) are frequently used as putative biomarkers in research on development, aging, and a myriad of neurological and psychiatric diseases.

Research using other modalities has typically applied the "frame displacement" calculated from timeseries realignment parameters to summarize motion during a scan (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). However, most structural imaging sequences do not provide a ready estimate of participant motion during acquisition. Recently, Tisdall and colleagues overcame this limitation by embedding volumetric navigators (T1-vNavs) in the T1-weighted volume acquisition, allowing for motion to be directly quantified in a manner akin to functional imaging time series (Tisdall et al., 2012). Subsequently, Reuter et al. (2015) used this T1-vNavs MPRAGE sequence to demonstrate in 12 healthy adults that motion during the T1 sequence was associated with spurious alterations of cortical thickness and cortical volume.

Despite the clear importance of such work, the vast majority of T1-weighted imaging sequences acquired to date lack T1-vNav technology, and thus cannot derive the quantitative assessment of motion used by Reuter et al. (2015). Accordingly, three important recent studies used motion during a functional imaging sequence acquired during the same scanning session as a proxy of in-scanner motion during the structural

scan (Alexander-Bloch et al., 2016; Pardoe et al., 2016; Savalia et al., 2017). This approach is based on the observation that participant motion tends to be highly correlated across acquisitions: individuals with high motion in one sequence tend to have high motion in other sequences (Pardoe et al., 2016; Yan et al., 2013). All three of these studies demonstrated that higher motion during a functional sequence acquired in the same session is associated with cortical thickness, even in those scans which passed manual quality assurance procedures. Furthermore, Salvia et al. (2017) demonstrated that the confounding influence of motion inflated the apparent effects of aging. While motion during a functional sequence is the most opportune proxy for motion during a structural scan, it nonetheless has important limitations. First, the ecological validity of the proxy is likely to vary with ordering effects, amount of time between scans, and other uncontrolled variables such as patient comfort. Another limitation is that it requires the addition of a functional scan, which may not have been acquired due to timing limitations or the study-specific hypotheses.

Based on this context, we had four related aims in the current study. First, we sought to use quantitative metrics to identify T1-weighted volumes considered unusable as determined by three expert manual raters. Second, we used these measures to quantify artifact in images considered usable by manual raters. Third, we described how quantitative metrics of T1-weighted image quality related to measures of brain structure (thickness, volume, density). Finally, we investigated the confounding impact of data quality on apparent estimates of brain development during adolescence. To accomplish these goals, we utilized summary measures of the T1-weigthed image provided by the Quality Assessment Protocol (QAP)(Shehzad et al., 2015), which is part of the Configurable Pipeline for Analysis of Connectomes (C-PAC)(Bellec et al., 2017). All measures were calculated for a large sample of T1-weighted volumes acquired as part of the Philadelphia Neurodevelopmental Cohort (PNC; n = 1601), which were independently manually rated by three trained experts. Multivariate models were constructed in a large training sample, and then tested in a completely independent validation sample. These model-based metrics were directly compared to motion as measured during functional sequences acquired in the same scanning session. Finally, limits of this approach were assessed by examining model performance in images acquired on a different scanner as part of separate study. As described in detail below, we found that measures derived directly from the T1-weighted volume provide highly reliable measures of image quality.

METHODS

Approach overview

Our overall goal was to evaluate quantitative measures of image quality directly from structural MRI volumes, and understand how such measures relate to commonly used measures of brain structure. This process included *seven* discrete tasks. First, three image analysts underwent rigorous training, and then independently rated each of the 1601 T1-weighted volumes from the PNC (Satterthwaite et al., 2014). Second, based on these manual ratings, we created a statistical model that sought to identify images that were unusable and should not be included in analyses of brain structure; we refer to this as the "inclusion" model. Third, we built a second model that could differentiate usable images

with some artifact from artifact-free images, ultimately using this information to yield a continuous measure of image quality; we refer to this as the "quantification" model. Fourth, we related the quantitative measures of image quality to commonly used measures of brain structure from two pipelines. Fifth, we examined how data quality can bias inference in studies of adolescent brain development. Sixth, we compared this approach to proxy measures estimated from motion during functional time series acquired during the same session. Seventh, we tested the limits of our strategy, and applied our models to a new dataset acquired on a different scanner.

Participants and structural data acquisition

Imaging data included 1601 scans from the PNC. All images were acquired on the same 3T Tim Trio scanner with a 32-channel head coil (Siemens: Erlangen, Germany) as previously described (Satterthwaite et al., 2014). Structural images were acquired using a magnetization-prepared, rapid-acquisition gradient-echo (MPRAGE) T1-weighted sequence (TR = 1810ms; TE = 3.51ms; T1 = 1100ms; FoV = 180 × 240mm; flip angle = 9°; GRAP PA factor = 2; BW/pixel = 130 Hz; resolution: 0.94mm x 0.94mm x 1.0mm; Acquisition time = 3:28). Prior to scanning, in order to acclimate participants to the MRI environment and to help subjects learn to remain still during the actual scanning session, a mock scanning session was conducted using a decommissioned MRI scanner and head coil. Mock scanning was accompanied by acoustic recordings of the noise produced by gradient coils for each scanning pulse sequence.

Manual rating procedure and rater training

Similar to prior work (Reuter et al., 2015; Savalia et al., 2017), all 1601 images were rated on quality using a 0-2 ordinal scale. Initial pilot testing indicated that using systems with finer categories (i.e., 4 or 5 rating classes) resulted in substantially diminished inter-rater reliability. In this framework, a "0" denoted images that suffer from gross artifacts and were considered unusable. In contrast, a "2" was assigned to images free from visible artifact. The intermediate "1" category was used for images with some artifact, but which still would be considered usable.

A rigorous process of training was used to ensure high inter-rater reliability (see **Figure 1**). First, anchors and exemplars for the three quality classes were agreed upon through consensus of 5 experts, including a board-certified neuroradiologist (JES), an MR physicist (MAE), a cognitive neuroscientist (DRR), an experienced image analyst (AR), and a neuropsychiatrist (TDS). Next, two of these experts (DRR and TDS) created a larger training sample by rating 100 images independently. Initial concordance was 93%; discrepancies were resolved through consensus, thus yielding a set of 100 images that were used to train three image analysts (KS, PV, JB) who served as the raters for the complete dataset. These three analysts were trained to >85% agreement in this dataset. This required two rounds of blind rating: during the first round, agreement with the expert consensus was 82%(JB), 57%(PV), and 82%(KS). Following further training, each rater re-rated this set of 100 images (presented in a different order, without identifiers), and achieved an accuracy of 91%(JB), 86%(PV), and 94%(KS). Having met reliability benchmarks, these three raters then independently rated all 1601 images of the PNC.

Rater concordance was evaluated using two measures: the weighted-κ statistic and polychoric correlations. These two measures provide complementary information: while

the weighted-κ assesses absolute rating agreement, the polychoric correlation assesses the ordering of the ratings. Variation amongst raters were assessed using a repeated effects ANOVA model. The relationship between manual rating and demographic variables (age and sex) was evaluated with a linear mixed effects model, implemented using the `lme4` (Bates et al., 2015) package in R. In this model and those described below, rater was modeled as a random effect to account for small but systematic differences between raters.

Construction of training and validation samples

Prior to model building, the complete collection of T1-weighted volumes was split into training and testing samples. The use of an independent validation set allowed us to select model parameters in the training sample, and then evaluate model performance in the validation sample without concern for over-fitting. Specifically, 1,067 images were used for training, and 534 were used during validation (see **Figure 2A**). In order to maintain a similar distribution of age, sex, and manual image quality rating across the two samples, we used the `caret` package in R (Kuhn et al., 2016;)(see **Table 1**).

Quantitative metrics of structural image quality

We evaluated the utility of an array of quantitative imaging measures included in QAP (see **Table 2**)(Shehzad et al., 2013). QAP version 1.0.3 utilizes FMRIB's Automated Segmentation Tool (FAST, Zhang et al., 2001) for image segmentation, which enables definition and quantification of quality metrics using an image's gray matter, white matter, and background voxels. Steps were taken to avoid the inclusion of neck and face tissue within the image's background for the calculation of all background metrics as previously described (Mortamet et al., 2009). Image kurtosis and skewness (Joanes and Gill, 1998) were also calculated for each tissue class and background using tools included in the `ANTsR` (Avants et al., 2016) and `psych` (Revelle, 2017) packages in R; these measures have been integrated into recently-released updates to QAP. In order to visualize the relationship between quantitative measures and manual quality rating, we plotted the mean value for each image quality metric versus the mean manual quality rating. Furthermore, we also calculated partial Spearman's correlations between manual ratings and automated metrics (while controlling for age, age squared, and sex). For these plots and analyses, we collapsed any image with at least one '0' rating into the '0' bin due to small cell size of other bins.

Creation of model to identify unusable images ("Inclusion" model)

Next, we sought to use the quantitative measures described above in a model, termed the "inclusion" model, to identify unusable images. We used a logistic mixed effects model (Bates et al., 2015), where rater was modeled as a random effect. This model sought to identify unusable images (i.e., rated "0") from usable images (i.e., rated "1" or "2"). The model was built in the training set (n = 1067). As each image was rated three times, there were in total 3201 observations in the training dataset.

Model training began with a simple mass-univariate model and then added features to create a multivariate model in a forward-stepwise manner. The first (base) variable in the multivariate model was defined as the variable with the best performing receiver operator curve (ROC) as measured by area under the curve (AUC) in the mass-univariate analyses. Each additional measure was added separately to this base model, and the AUC

was re-calculated. The best performing feature was selected, and this process was repeated. At each step, in order to determine whether an additional model parameter provided significantly improved classification, we calculated the Delong statistic, which tests for a significant increase in AUC (DeLong et al., 1988). Model building was terminated when no significant increase in AUC was found for any variable.

Performance in the training set was assessed by AUC, accuracy, sensitivity, and specificity. Next, the exact parameters and classification threshold criterion from the training set were applied to the validation set, and the same outcome measures were evaluated in this independent dataset. For the validation data (n = 534), we collapsed manual ratings to derive a single label per image (rather than a label for each rater), where any image with at least one '0' rating was considered unusable (n = 16 images).

Creation of model to identify image artifacts in usable data ("Quantification model")

While the "inclusion" model identified unusable data (rated "0"), we next aimed to identify more subtle variations in image quality, which we call the "quantification" model. Identical procedures were used as described above, except that the task of the model was to differentiate images rated "1" and "2"; images rated "0" were not included in this analysis. The training sample for the quantification model was comprised of 1,034 images, with 269 images with some artifact (rated "1"). As above, to derive a single label per image in the validation dataset, we collapsed information across raters and assigned the modal manual rating.

Notably, while the quantification model remained a logistic mixed model, it also yielded a predicted quality value ($\hat{y}$) that was continuously distributed. In this model, a larger $\hat{y}$ value (log-odds) denoted higher data quality and diminished artifact. As described below, we next compared the relationship between the continuous "quantification" model prediction ($\hat{y}$) and commonly-used morphometric measures derived from the structural image.

Structural image processing

We evaluated multiple measures including cortical volume, thickness and gray matter density. To parcellate the brain into anatomically-defined regions, we used an advanced multi-atlas labeling approach. Specifically, 24 young adult T1-weighted volumes from the OASIS data set (Marcus et al., 2007) that were manually labeled by Neuromorphometrics, Inc. were registered to each subject's T1-weighted volume using the top-performing SyN diffeomorphic registration (Avants et al., 2011a; Klein et al., 2010). These label sets were synthesized into a final parcellation using joint label fusion (Wang et al., 2013). Volume was determined for each parcel using the intersection between the parcel created and the intersection of a prior driven gray matter cortical segmentation from the ANTs cortical thickness pipeline as described below. Mean thickness and density were calculated within each parcel as described below.

Cortical thickness was quantified using tools included in ANTs (Tustison et al., 2010). To avoid registration bias and maximize sensitivity to detect regional effects that can be impacted by registration error, a custom adolescent template and tissue priors were created using data from 140 PNC participants, balanced for age and sex. Structural images were then processed and registered to this custom template using the ANTs cortical thickness pipeline (Tustison et al., 2014). This procedure includes brain extraction, N4 bias

field correction (Tustison et al., 2010), Atropos tissue segmentation (Avants et al., 2011b), SyN diffeomorphic registration method (Avants et al., 2011a; Klein et al., 2010), and diffeomorphic registration-based cortical thickness (DiReCT) estimation in volumetric space (Das et al., 2009). One large-scale evaluation study has shown that this procedure for estimating cortical thickness is more sensitive to individual age and sex differences over the lifespan than FreeSurfer (Tustison et al., 2014). However, in order to ensure that results were not specific to this processing pipeline, we also calculated regional volume and thickness using FreeSurfer v5.3 (Fischl and Dale, 2000).

Finally, gray matter density was calculated using Atropos (Avants et al., 2011b), using an iterative segmentation procedure that is initialized using 3-class K-means segmentation. This procedure produces both a discrete 3-class hard segmentation as well as a probabilistic gray matter density map (soft segmentation) for each subject. Gray matter density was calculated within the intersection of this 3-class segmentation and the subject's volumetric parcellation (Gennatas, et al., under review).

Relationship of quality estimates to measures of brain structure

Initial comparisons between data quality and measurements of brain structure were completed on a global level, including mean cortical thickness, total cortical volume, and mean gray matter density. This was performed in useable images (i.e., those that did not receive a "0" rating by any of the three manual raters). Both mean thickness and total volume were evaluated for ANTs and FreeSurfer; density was examined using ANTs only. In order to relate these global image estimates to our quantification model's outcome, a partial Spearman correlation was calculated; age, age squared, and sex were included as covariates in the calculation of these partial correlations.

Following this evaluation of global measures of brain structure, we examined associations between data quality and regional brain structure. Specifically, we used linear regression to examine the association between image quality as summarized by the quantification model and regional estimates of cortical thickness, cortical volume, and gray matter density calculated using the ANTs pipelines. Age, age squared, and sex were included in these regression models as covariates. Multiple comparisons were controlled for using the False Discovery Rate (FDR; q = 0.05).

Impact of data quality on estimates of brain development

The analysis described above revealed substantial relationships between data quality and measures of brain structure. We next examined how data quality may bias estimates of brain development. Specifically, we conducted mediation analyses, to determine to what degree data quality (as summarized by the $\hat{y}$ of the quantification model) might mediate the apparent relationship between age and brain structure. A Sobel's *z*-score (Sobel, 1982) was calculated for each cortical region for thickness, volume, and gray matter density. Sobel's *z*-score estimation was implemented in the `bda` package in R (Wang, 2015). Multiple comparisons were controlled for using FDR as above (q = 0.05).

Comparison to motion in functional scans acquired in the same session

Three recent reports demonstrated that motion in functional sequences acquired during the same scanning session acts as an effective proxy for structural image quality (Alexander-Bloch et al., 2016; Pardoe et al., 2016; Savalia et al., 2017). Accordingly, derived

T1 based quantitative metrics of image quality were then compared to head motion estimated from functional sequences acquired later during the same scanning session. These additional sequences included a pseudo-continuous arterial spin labeled (PCASL) perfusion scan, two task fMRI scans (tfMRI 1 & tfMRI 2), and one resting functional connectivity scan (rsfMRI) (Satterthwaite et al., 2014, 2016). The motion in each of these sequences was summarized as the Frame Displacement (FD), estimated using the average root mean square displacement as calculated by FSL's MCFLIRT (Jenkinson et al., 2002). As not all individuals completed all functional sequences in the PNC protocol, this analysis was restricted to 804 subjects in the training sample and 424 subjects in the validation sample for whom complete data was available. For each functional sequence, the partial Spearman correlation value between the FD of each sequence and each global measure of brain structure were calculated as above; age, age squared, and sex were included as covariates as prior. In order to compare the magnitude of these relationships with the quantification model, the absolute value of these correlations was plotted.

Generalizability of T1 based quantitative metrics to data acquired on a different scanner

In order to assess the limits of our strategy and generalizability to images acquired on different scanning hardware, we tested our models on T1-weighted volume data from another study (Roalf et al., 2015). In this study, T1-weighted volumes were collected on a different 3T Tim Trio scanner, using an 8-channel head coil with the following acquisition parameters: TR = 1680ms; TE = 4.67ms; T1 = 1100ms; FoV = 180 × 240mm; flip angle = 15°; bandwith/pixel = 150Hz; resolution: 0.94mm x 0.94mm x 1.0mm; acquisition time = 5:00 (Roalf et al., 2015). Instead of a large sample of youth, this study compared healthy adults and those with a family history of schizophrenia.

Data from this study was rated by each of the same three raters as described above, and processed using identical procedures. The inclusion and quantification models developed on the PNC were applied to this new dataset, and assessed according to their AUC, accuracy, sensitivity, and specificity. Finally, new models were developed on this dataset using identical procedures to test performance within this novel data set.

RESULTS

Highly trained manual raters achieve good concordance

Although there were significant differences among raters (training: $F[2, 3198] = 39.65$, $p<.0001$; validation: $F[2, 1599] = 17.74$, $p<.0001$) a post hoc analysis found that raters never disagreed by more than one quality class. The weighted kappa statistic indicated that three raters achieved good concordance (**Figure 2B**) in both the training (mean weighted-$\kappa = 0.64$) and validation dataset (mean weighted-$\kappa = 0.68$). Additionally, polychoric correlations (**Figure 2C**), indicated very high correlation amongst raters in both the training (mean $r = 0.93$) and validation samples (mean $r = 0.94$).

Manual quality rating varies by age and sex

As expected, age and sex were associated with manual quality rating (**Figure 3**). In both the training and validation dataset, males had a trend towards more in-scanner head motion than females (training: $\beta = .04$, $t = 3.73$, $p < 0.0001$; validation: $\beta = .08$, $t = 4.66$, p <

0.0001; **Figure 3A**). Similarly, younger age was associated with lower quality (training: β = .02, t = 10.50, p < 0.0001; validation: β = .03, t = 12.68, p < 0.0001; **Figure 3B**), emphasizing the degree to which data quality could potentially confound studies of brain development.

Quantitative measures of image quality align with manual rating

Comparing quantitative image-based measures of quality revealed marked heterogeneity (**Figure 4**). Among quantitative measures, the strongest associations with manual rating were observed for background kurtosis and skewness, white matter skewness, and Qi1. While the relationship between manual rating and quantitative measures was strong and mostly linear for some measures (e.g., Qi1), other measures demonstrated a nonlinear relationship or only weak associations. For example, unusable images (rated "0") had disproportionately lower background skewness and kurtosis compared to usable images (rated "1" or "2"). Relationships between manual rating and quantitative metrics were highly consistent across the training and validation samples.

Structural imaging quality metrics allow identification of unusable scans

We next used the T1 quantitative metrics to build an "inclusion" model that discriminated unusable images (rated "0") and usable images (rated "1" or "2"). We built this model using a stepwise process of variable selection in the training sample, and testing this model in the validation sample. As an initial step, we conducted mass-univariate analyses that evaluated the ability of each quantitative imaging metric to identify a usable image (**Figure 5A**). Background kurtosis demonstrated the best performance (AUC = 0.88) and was thus the first variable included in the model. Next, we examined whether adding an additional variable improved performance above and beyond background kurtosis. Notably, adding background skewness to the model significantly improved model performance (AUC = 0.95, Delong p < 0.0001). However, adding a third variable to this model did not significantly improve ROC performance. Thus, the final inclusion model was:

$$Useability\ (0\ vs\ !\ 0) \sim background\ kurtosis + background\ skewness + (1|rater)$$

In this model "1|rater" indicates the inclusion of manual rater as a random factor.

This model performed well within the training data set (AUC = 0.95, classification accuracy = 0.83; **Figure 5B**). Furthermore, when this model was applied to the validation set, performance remained high (AUC = 0.94, classification accuracy = 0.80; **Figure 5C**). In both the training and validation dataset, this model had moderately high sensitivity, and very high specificity (see **Table 2**).

Image artifact can be quantified using T1-based measures

Having established that unusable images could be identified by quantitative metrics, we next sought to build a model that could quantify artifact in usable images. Specifically, this model sought to discriminate images rated "1" from those rated "2". As expected for this (more difficult) task, mass-univariate analyses revealed performance that was less robust than when these metrics were used to identify unusable images. Qi1 performed best (AUC = 0.70), followed closely by white matter skewness (AUC = 0.70) and background kurtosis (AUC = 0.69; **Figure 6A**). When the model was expanded in a stepwise manner, we found that significant increases in AUC stopped after 7 steps, yielding a model with 8

variables. The quantification model was:

$$Artifact\ (1\ vs. 2) \sim Qi1 + white\ matter\ skewness + CNR + background\ kurtosis + EFC + background\ skewness + FBER + SNR + (1|rater)$$

This model performed well in both the training dataset where variable selection was performed (AUC = 0.80, classification accuracy = 0.73; **Figure 6B**), and also in the independent validation set (AUC = 0.92, classification accuracy = 0.73; **Figure 6C**). In both the training and validation dataset, this model had moderately high sensitivity, and very high specificity (see **Table 3**).

Integrated quantitative image quality metrics are related to measures of brain structure

We next evaluated the degree to which quantitative measures of image quality were related to commonly used measures of brain structure. These measures included cortical thickness, cortical volume, and gray matter density; thickness and volume were evaluated for both ANTs and FreeSurfer pipelines. For most measures derived from the ANTs pipeline, the outcome of the quantification model and the average manual rating demonstrated similar correlations in training and validation samples (**Figure 7**). As expected and in line with prior reports, higher quality data was associated with thicker cortex on average. Total cortical volume was less impacted by artifact, and only had a nominal relationship with manual rating in the training dataset. However, in both the training and validation dataset, gray matter density displayed a very strong relationship with both manual and quantitative measures of artifact, with higher quality being associated with higher density.

To ascertain if these results were specific to the processing pipelines utilized, we re-ran the above analyses using the FreeSurfer software suite. (Density was not evaluated as that is not a standard output of FreeSurfer). Broadly, convergent results were revealed. However, the association between data quality and cortical thickness was substantially stronger using data from FreeSurfer in both the training and validation samples. Volumetric measures derived from FreeSurfer also appeared to be substantially more sensitive to image artifact than those derived from multi-atlas labeling procedure implemented in ANTs. However, the quantification model diverged for this measure from manual ratings and was not significantly related to total cortical volume as calculated using FreeSurfer.

To further understand such global associations with data quality, we conducted mass-univariate analyses evaluating the relationship between data quality (as summarized by the $\hat{y}$ from the quantification model) with regional thickness, volume, and density calculated using ANTs. For cortical thickness, of the 98 regions evaluated, there was an FDR-corrected relationship with data quality in 75% of cortical regions (**Figure 8A**) in the training dataset and 62% in the validation dataset (**Figure 8B**). However, the directionality of this association was regionally heterogeneous. In sensorimotor and temporal regions, higher data quality was associated with thicker cortex. In contrast, regions including the anterior cingulate, dorsolateral prefrontal cortex, and superior parietal cortex, higher data quality was associated with thinner cortex.

Regional volume as measured by ANTs was somewhat less impacted by data quality, with only 11% of ROIs demonstrating a significant association with image quality (**Figure**

8C) in the training and 15% in the validation data set (**Figure 8D**). Finally, gray matter density was heavily impacted by artifact, with every region demonstrating an FDR-significant association with image quality in both the training and validation samples (**Figure 8D** & **E**). For all regions, higher data quality was associated with higher gray matter density.

Data quality systematically biases estimates of structural brain development

Having established that data quality as measured directly from features of the T1 image is related to commonly used measures of brain structure, we next evaluated the degree to which data quality might alter inference regarding brain development in youth. This is of particular concern as age is systematically related to data quality (i.e., **Figure 3A**). Accordingly, we conducted mediation analyses to examine the degree to which data quality might mediate the relationship between age and brain structure. As expected from regionally heterogeneous effects of data quality on cortical thickness (see **Figure 8A**), data quality had a bidirectional bias on estimates of development (**Figure 9A)**. For most regions (shown in red), the relationship with data quality resulted in a masking of developmental effects, and age effects became more significant when controlling for data quality. However, in several regions primarily located in cingulate cortex (shown in blue), data quality had the opposite effect, and inflated apparent developmental effects. Results were highly concordant in the training and validation datasets.

In contrast, data quality significantly mediated the development of gray matter density in a consistent manner across all regions (**Figure 9E** & **F**). As data quality reduced gray matter density, and density increases with age, artifact in the T1-weighted volume inflated apparent associations with age, and controlling for data quality resulted in diminished significance of age effects.

Motion from functional sequences align variably with manual rating and quantitative measures

Based on prior reports that motion in functional sequences acquired in the same scanning session can provide a useful proxy of structural image quality, we compared quantitative measures of structural image quality to frame displacement from the four functional scans acquired as part of the PNC. Mean motion in each functional sequence increased as the scanning session progressed in both the training and validation (**Figure 10 A** & **B**). Next, we compared these proxy measures to global measures of brain structure (**Figure 10C** & **D**). In general, the magnitude of the relationship between motion during each functional sequence and brain structure was similar to that of the quantification model. As might be expected, functional sequences that were collected at the end of the scanning session (and thus were farther away in time from the T1 scan) tended to have a weaker relationship with measures of brain structure. This was most obvious for the resting state sequence, which was acquired last. As noted above, quantitative measures of image quality did not align well with total cortical volume.

Model of structural image quality does not to generalize to data acquired on a different scanner

As a final step, we sought to establish limits in the generalizability of these methods, and evaluated whether the models trained on a single scanner were generalizable across

hardware platforms. Models trained on the PNC dataset failed to accurately identify unusable images (inclusion model: AUC = 0.48) or artifact-contaminated images (quantification model: AUC = 0.58). In contrast, models trained on data from the new scanner performed quite well for both identifying unusable images (inclusion model AUC = 0.97) and usable images impacted by artifact (quantification model AUC = 0.84) within the novel dataset.

DISCUSSION

In this paper, we demonstrate that summary measures derived from a T1-weighted volume can replicate results from visual inspection with a high accuracy. These measures were comparable to proxy measures derived from functional sequences acquired in the same scanning session. Furthermore, we demonstrate that image-based measures of data quality show differential relationships to several common measures of brain structure, and systematically alter estimates of brain development in youth. However, we caution that our classifiers did not generalize across scanning platforms, emphasizing the need for additional research to identify robust automated classifiers of data quality. This effort is significantly assisted by openly shared neuroimaging datasets and tools such as the PNC and the QAP software package.

Manual raters can achieve a high level of concordance in a large-scale sample

It is increasingly recognized that data quality may be the primary confound in brain imaging studies of individual difference, lifespan development, or clinical populations (Ciric et al., 2017; Power et al., 2015). In-scanner motion is usually the single biggest determinant of data quality, especially in individuals who are young, elderly, or ill. While summary measures of motion can be easily derived from the realignment parameters of functional time series, motion cannot be easily estimated for most existing structural imaging data. Volumetric navigators (Tisdall et al., 2012) provide an extremely useful solution to this problem (Tisdall et al., 2016), but such sequences have not been used for the vast majority of already-collected imaging data, which represents a huge societal investment. Due to lack of ground truth, one of the first challenges for any study attempting to estimate image-derived measures of data quality for structural images is to create manual ratings, which are necessary to train subsequent quantitative models. This problem is quite analogous to studies of psychiatric or neurologic illness, where several clinicians evaluate information from a patient and attempt to come to a consensus.

With limited training utilities available, we pursued an approach analogous to established procedures for training on clinical interviews and rating scales (Forbes et al., 2010; Kaufman et al., 1997). Specifically, a panel of experts originally created a small set of anchors. Notably, while we originally used a rating system with 5 levels similar to that used in one recent study (Pardoe et al., 2016), we found that even highly trained experts could not reach a high level of concordance. Accordingly, we limited the quality rating to three levels, more akin to previous efforts (Reuter et al., 2015; Savalia et al., 2017). A larger training set of 100 images were then rated according to these anchors by two faculty experts, and were then used to train three experienced staff members to >85% accuracy. After this degree of reliability was established, the full set of images was evaluated.

Following such training, concordance remained relatively good in both the training and validation samples. The pairwise correlation between raters was even higher, reflecting that when raters were not concordant it was usually due to a small but significant between-individual rater bias. To capture such bias in all models of data quality, a mixed-effect framework was used, where rater was modeled as a random effect.

Image-based summary measures accurately predict image quality

Having established a reliable set of manual ratings, we next derived quantitative measures of data quality using summary statistics from the structural image alone. These measures were derived from the Quality Assurance Pipeline (QAP)(Shehzad et al., 2015) included in the Configurable Pipeline for Analysis of Connectomes (C-PAC)(Sikka et al., 2013). Prior studies have shown that utilizing a multivariate approach with a wide array of imaging features provides enhanced accuracy for artifact detection (Pizarro et al., 2016).

Using these measures, our first task was to identify images that were so corrupted by artifact as to be unusable. This is a common first step in any imaging study. We found that these images could be accurately identified using very few model parameters. Specifically, background skewness and background kurtosis were able to identify unusable images with an AUC of 0.95 in the training set, and an AUC of 0.94 in an independent validation dataset. These measures reflect the change in the shape of the distribution of background voxels, as artifacts introduce more high intensity voxels into the background.

Having identified images that are unusable we removed them from the analysis and created a model trained to differentiate images without notable artifact and useable images where some artifact was present. This model yields both a dichotomous classification and a dimensional score that could be used as a covariate in group-level analyses, akin to the use of frame displacement in fMRI studies. As anticipated, this was a more difficult task: using 8 model parameters, AUC in the training set was 0.80, with a validation AUC of 0.92. The most predictive feature in this model was Qi1, which has previously been examined as a quantitative measure of image quality in large scale studies (Mortamet et al., 2009). While Qi1 was the most discriminative feature, it alone produced an AUC of only 0.70. However, with of 7 additional features including skewness, kurtosis, CNR, SNR, FBER, and EFC raised model performance significantly (AUC = 0.80). Together, results of both models demonstrate that summary measures from the structural image itself can be used to derive a quantitative estimate of data quality. This information can be used to identify and exclude unusable images, and potentially as a nuisance covariate in group-level analyses.

Measures of brain structure are differentially impacted by data quality

As noted above, the quantification model yielded not only a classification, but also a dimensional index of image quality. This fully continuous quality score was next related to commonly-used measures of brain structure. Previous work has shown that both cortical thickness and volume is systematically biased by in-scanner motion, whether quantified by manual rating (Pardoe et al., 2016; Savalia et al., 2017), motion estimated from functional sequences acquired in the same scanning session (Alexander-Bloch et al., 2016; Pardoe et al., 2016; Savalia et al., 2017), or volumetric navigators embedded in the T1 sequence (Reuter et al., 2015). Here, we demonstrate that the index of image quality derived directly from the structural image itself shows a similar relationship. Relationships between measures of brain structure and our quantitative index were nearly as strong as the

associations with the average of three expert raters. Realistically, most studies will not devote the resources to training three raters to a high degree of concordance, and require each rater to examine every structural image.

By evaluating a broad set of structural measurements, we demonstrated that the relationship with data quality is heterogeneous, and some measures are much more sensitive to noise than others. As reported in other studies, we found that higher data quality was globally associated with thicker cortex. There was substantial variability across pipelines, and cortical thickness as measured by ANTs was less impacted by data quality than the estimates provided by FreeSurfer. Furthermore, this association had notable regional heterogeneity. In somatomotor, temporal, and many frontal regions, higher data quality was associated with greater thickness. However, in other regions within the dorsolateral prefrontal cortex, superior parietal cortex, and cingulate cortex, higher data quality was associated with thinner estimated cortical thickness. These results are strikingly convergent with prior reports including that of Reuter et al. (2015), where motion was measured directly from the scan using embedded volumetric navigators. It should be noted that regional heterogeneity of motion effects to some degree show correspondence with large scale functional networks (such as the fronto-parietal control network and cingulo-opercular networks (Dosenbach et al., 2007; Vincent et al., 2008)). Thus, regional heterogeneity in the effect of data quality on cortical thickness may masquerade as results with interpretable functional neuroanatomy.

Of the measures evaluated, gray matter volume was the most robust to changes in image quality. Total gray matter volume as measured using ANTs was not significantly related to image quality. However, there was evidence of a substantial effect of processing pipeline, and with gray matter volume as measured by FreeSurfer being more strongly related to data quality. Such variation may be due to the multi-atlas labelling procedure employed, which has previously been found to be highly accurate in rigorous analysis grand challenges (Landman and Warfield, 2012). Regardless of the pipeline used, quantitative measures of image quality were not significantly related to mean brain volume, suggesting that this approach may have differential utility depending on the primary measure of interest.

Beyond volume and cortical thickness, we also evaluated gray matter density, a measure of brain structure derived from image segmentation. Gray matter density is a commonly used metric in voxel-based morphometry studies, with applications to neuropsychiatric and neurodegenerative illnesses (Gogtay et al., 2004; Meijer et al., 2011). Our analyses revealed a marked impact of data quality on gray matter density. Taken together, these results emphasize that data quality as estimated directly from the structural image is associated with bias in commonly-used measures of brain structure, and suggest that FreeSurfer may be somewhat more sensitive to such noise than ANTs. Finally, we provide new data showing that gray matter density may be particularly sensitive to such bias.

Measures of motion from functional sequences show a variable relationship to structural image quality

One recently proposed approach for deriving estimates of data quality from structural images is to use motion estimated from a functional time series acquired within the same session as a proxy. Several prior reports have shown that this is a fruitful

approach (Alexander-Bloch et al., 2016; Pardoe et al., 2016; Savalia et al., 2017), demonstrating associations between this proxy measure of data quality and both cortical thickness and volume. One clear limitation of this approach is that it requires a functional scan to be acquired in the same scanning session. We document a second limitation, which is that the relationship between data quality as estimated from the functional sequence and the structural image may vary considerably based on the order of sequence acquisition. Thus, the ASL scan acquired immediately after the T1 image in our study showed a stronger relationship to measures of brain structure, which were similar in magnitude to the manual rating and image-based quality index. However, the resting-state scan acquired at the end of our scanning session showed markedly attenuated relationships in both the training and validation datasets, most likely due to increased variation and amplitude of movement as the scanning session progressed. These data should caution investigators that proxy measures of motion from functional sequences are not uniformly informative, and that sequences acquired closer in time to the structural image are more likely to provide useful proxy data.

Data quality biases estimates of structural brain development in youth

Accurate measurement of cortical thickness, volume and gray matter density are critical to understanding typical and atypical trajectories of the developing brain. The extant literature indicates robust age-related cortical thinning, reductions in regional gray matter volumes, and increases in gray matter density in adolescence (Courchesne et al., 2000; Gennatas et al., n.d.; Giedd et al., 1999; Giedd, 2004; Gogtay et al., 2004; Jernigan et al., 1991; Kennedy et al., 2002; Lenroot and Giedd, 2006; Pfefferbaum et al., 1994; Reiss et al., 1996; Shaw et al., 2008; Sowell et al., 2001, 2003, 2004). Moreover, there are regional-specific patterns of cortical maturation throughout development, with delayed maturation of higher-order association cortex (Giedd et al., 1999; Giedd, 2004; Gogtay et al., 2004; Shaw et al., 2008; Sowell et al., 2004; Tamnes et al., 2010). Yet, the robustness of these results is predicated upon high data quality. While most of these studies use validated methods to reduce in-scanner head motion during acquisition, few if any systematically evaluated or controlled for data quality. Importantly, several recent reports indicate significant relationships between age and in-scanner head motion in a variety of MRI protocols (Power et al., 2012; Satterthwaite et al., 2012; Roalf et al., 2016). Thus, our findings of reduced data quality in some T1-weighted volumes were not surprising. To see if previously reported developmental trends are resilient to the negative impact of data quality we performed region-wise mediation analyses. Age-related changes in regional volume were largely resilient to data quality. However, associations between age and both cortical thickness and density were significantly altered. Age-related changes in regional cortical thickness showed bidirectional, regionally heterogeneous bias when data quality was considered. Several regions in frontal, temporal, parietal and occipital cortices showed more prominent developmental effects once T1 data quality was considered. In contrast, medial regions, such as the anterior cingulate, show less prominent associations with age after controlling for data quality. Mediation analyses of age and gray matter density indicated that not accounting for data quality results in inflated apparent developmental associations.

Limitations

Despite the accuracy of our image-based index of data quality, application of this model to an external dataset acquired on a different scanner provides evidence of clear limitations. While the model trained on a single scanner retained superb performance, when faced with previously unseen data acquired in the same way from the same scanner, performance was poor for data acquired on a different scanning system. This is likely due to differences in sensitivity of the head coil, which is known to influence many measures of image intensity (Mortamet et al., 2009); differences in the sample composition (adults vs. youth) may have contributed as well. Constructing a new model with data from the new scanner restored the classification model to excellent performance, showing that approach is effective as long as it is trained and tested upon data from the same scanner. These results therefore emphasize that our statistical models of quality indices constructed from summary statistics calculated directly from the T1 image are hardware-specific, and do not generalize across scanners. Future work should evaluate which sequence and hardware variations specifically lead to such limitations, and evaluate heterogeneity amongst scanners, coils, and software. Furthermore, the degree to which such measures could provide useful information regarding the quality of other structural imaging sequences (e.g., T2-weighted images) remains unknown. Finally, there is clear need for additional research to evaluate whether a different modeling approach (perhaps using machine learning techniques) could produce a more generalizable classifier using similar input features.

Conclusions

The present work demonstrates that data quality can be estimated directly from structural images that lack volumetric navigators. Such image-based indices of data quality can be used to exclude unusable images in a reproducible fashion. Furthermore, these continuous measures of image quality have the potential to be used as covariates in group-level analyses of structural imaging data. The ability to derive a measure of data quality directly from the structural image obviates the need for use of proxy measures from functional sequences, which we show have heterogeneous utility based on their proximity of acquisition to the structural image in the scanning session.

More broadly, the present data emphasize the degree to which data quality should be appreciated as an important confound in structural imaging studies, and that this varies based on the measure and analysis pipeline used. We present data establishing that gray matter density is severely impacted by data quality. Cortical thickness is substantially impacted as well, while volume shows more attenuated relationships with image quality. Furthermore, thickness and volume estimates derived from FreeSurfer appear to be more impacted by data quality than pipelines that use ANTs.

Given such data, investigators are encouraged to report measures of data quality for all structural imaging studies that evaluate individual or group differences. This is particularly relevant for studies where data quality is likely to be systematically related to the primary subject-level variable of interest, such as age, clinical group status, or disease severity. We provide one such example, with results that show that data quality can systematically bias estimates of structural development during youth. While it is now common practice to report summary measures of motion and image quality for fMRI research, it is less common for studies using T1-weighted imaging. The present results

underscore a need for transparent reporting of such data. We urge investigators to report associations between data quality and both subject level variables of interest (e.g., age, group) as well as the primary imaging measure evaluated.

As noted above, the lack of generalizability of models trained on one scanner to data acquired on a different scanner will limit groups from using the exact model generated for the PNC and described here. However, given that MRI scanners represent a substantial financial investment (of several million dollars), it is not unreasonable to anticipate that such a model might be created for data acquired on each platform, and applied to the many studies collected using that hardware. Moving forward, volumetric navigators may obviate the need for such an approach, but we anticipate that the strategy outlined here will likely prove to be useful for the massive amount of structural imaging data that has already been collected at great effort and cost.

ACKNOWLEDGEMENTS

We thank the acquisition and recruitment team, including Karthik Prabhakaran and Jeff Valdez. Thanks to Chad Jackson for data management and systems support. Supported by grants from the National Institute of Mental Health: R01MH107703 (TDS) and R01MH107235 (RCG). The PNC was funded through NIMH RC2 grants MH089983 and MH089924 (REG). Additional support was provided by K01MH102609 (DRR), P50MH096891 (REG), R01NS085211 (RTS), K01ES026840 (JES), and the Dowshen Program for Neuroscience. Support for developing statistical analyses (RTS & TDS) was provided by a seed grant by the Center for Biomedical Computing and Image Analysis (CBICA) at Penn.

REFERENCES

Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J., Raznahan, A., 2016. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. Hum. Brain Mapp. 37, 2385–2397. doi:10.1002/hbm.23180

Atkinson, D., Hill, D.L., Stoyle, P.N., Summers, P.E., 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. doi:10.1109/42.650886

Avants, B.B., Kandel, B.M., Duda, J.T., Cook, P.A., Tustison, N.J., Shrinidhi, K., 2016. ANTsR: ANTs in R: quantification tools for biomedical images.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011a. A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration. NeuroImage 54, 2033–2044. doi:10.1016/j.neuroimage.2010.09.025

Avants, B.B., Tustison, N.J., Wu, J., Cook, P.A., Gee, J.C., 2011b. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9, 381–400. doi:10.1007/s12021-011-9109-y

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4 | Bates | Journal of Statistical Software 67, 1–48. doi:10.18637/jss.v067.i01

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D.S., Craddock, R.C., 2017. The Neuro Bureau ADHD-200 Preprocessed repository. NeuroImage, Data Sharing Part II 144, Part B, 275–286. doi:10.1016/j.neuroimage.2016.06.034

Bellon, E., Haacke, E., Coleman, P., Sacco, D., Steiger, D., Gangarosa, R., 1986. MR artifacts: a review. Am. J. Roentgenol. 147, 1271–1281. doi:10.2214/ajr.147.6.1271

Chen, J., Liu, J., Calhoun, V.D., Arias-Vasquez, A., Zwiers, M.P., Gupta, C.N., Franke, B., Turner, J.A., 2014. Exploration of scanning effects in multi-site structural MRI studies. J. Neurosci. Methods 230, 37–50. doi:10.1016/j.jneumeth.2014.04.023

Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., Gur, R.C., Gur, R.E., Bassett, D.S., Satterthwaite, T.D., 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. NeuroImage. doi:10.1016/j.neuroimage.2017.03.020

Courchesne, E., Chisum, H.J., Townsend, J., Cowles, A., Covington, J., Egaas, B., Harwood, M., Hinds, S., Press, G.A., 2000. Normal brain development and aging: quantitative analysis at in vivo MR imaging in healthy volunteers. Radiology 216, 672–682. doi:10.1148/radiology.216.3.r00au37672

Das, S.R., Avants, B.B., Grossman, M., Gee, J.C., 2009. Registration based cortical thickness measurement. NeuroImage 45, 867–879. doi:10.1016/j.neuroimage.2008.12.016

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Dosenbach, N.U.F., Fair, D.A., Miezin, F.M., Cohen, A.L., Wenger, K.K., Dosenbach, R.A.T., Fox, M.D., Snyder, A.Z., Vincent, J.L., Raichle, M.E., Schlaggar, B.L., Petersen, S.E., 2007. Distinct brain networks for adaptive and stable task control in humans. Proc. Natl. Acad. Sci. U. S. A. 104, 11073–11078. doi:10.1073/pnas.0704320104

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. U. S. A. 97, 11050–11055. doi:10.1073/pnas.200033797

Forbes, C., Blanchard, J.J., Bennett, M., Horan, W.P., Kring, A., Gur, R., 2010. Initial Development and Preliminary Validation of a New Negative Symptom Measure: The Clinical Assessment Interview for Negative Symptoms (CAINS). Schizophr. Res. 124, 36–42. doi:10.1016/j.schres.2010.08.039

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., and Turner, R. (1996). Movement-related effects in fMRI time-series. Magn Reson Med 35, 346-355.

Friedman, L., Glover, G.H., Krenz, D., Magnotta, V., FIRST BIRN, 2006. Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. NeuroImage 32, 1656–1668. doi:10.1016/j.neuroimage.2006.03.062

Gennatas, E.D., Avants, B.B., Wolf, D.H., Satterthwaite, T.D., Ruparel, K., Ciric, R., Hakon, H., Gur, R.E., Gur, R.C., n.d. Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to young adulthood.

Giedd, J.N., 2004. Structural magnetic resonance imaging of the adolescent brain. Ann. N. Y. Acad. Sci. 1021, 77–85. doi:10.1196/annals.1308.009

Giedd, J.N., Blumenthal, J., Jeffries, N.O., Castellanos, F.X., Liu, H., Zijdenbos, A., Paus, T., Evans, A.C., Rapoport, J.L., 1999. Brain development during childhood and adolescence: a longitudinal MRI study. Nat. Neurosci. 2, 861–863. doi:10.1038/13158

Gogtay, N., Giedd, J.N., Lusk, L., Hayashi, K.M., Greenstein, D., Vaituzis, A.C., Nugent, T.F., Herman, D.H., Clasen, L.S., Toga, A.W., Rapoport, J.L., Thompson, P.M., 2004. Dynamic mapping of human cortical development during childhood through early adulthood. Proc. Natl. Acad. Sci. U. S. A. 101, 8174–8179. doi:10.1073/pnas.0402680101

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.

Jernigan, T.L., Trauner, D.A., Hesselink, J.R., Tallal, P.A., 1991. Maturation of human cerebrum observed in vivo during adolescence. Brain J. Neurol. 114 (Pt 5), 2037–2049.

Joanes, D.N., Gill, C.A., 1998. Comparing measures of sample skewness and kurtosis. J. R. Stat. Soc. Ser. Stat. 47, 183–189. doi:10.1111/1467-9884.00122

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., Ryan, N., 1997. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. J. Am. Acad. Child Adolesc. Psychiatry 36, 980–988. doi:10.1097/00004583-199707000-00021

Kennedy, D., N., Makris, N., Herbert, M.R., Takahashi, T., Caviness Jr., V.S., 2002. Basic principles of MRI and morphometry studies of human brain development. Dev. Sci. 5, 268–278.

Klein, A., Ghosh, S.S., Avants, B., Yeo, B.T.T., Fischl, B., Ardekani, B., Gee, J.C., Mann, J.J., Parsey, R.V., 2010. Evaluation of volume-based and surface-based brain image registration methods. NeuroImage 51, 214–220. doi:10.1016/j.neuroimage.2010.01.091

Kuhn, M., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, and T., 2016. caret: Classification and Regression Training.

Landman, B.A., Warfield, S.K., 2012. MICCAI 2012 workshop on multi-atlas labeling," in Medical Image Computing and Computer Assisted Intervention Conference 2012: MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling Challenge Results.

Lenroot, R.K., Giedd, J.N., 2006. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. Neurosci. Biobehav. Rev. 30, 718–729. doi:10.1016/j.neubiorev.2006.06.001

Magnotta, V.A., Friedman, L., 2006. Measurement of Signal-to-Noise and Contrast-to-Noise in the fBIRN Multicenter Imaging Study. J. Digit. Imaging 19, 140–147. doi:10.1007/s10278-006-0264-x

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19, 1498–1507. doi:10.1162/jocn.2007.19.9.1498

Meijer, J.H., Schmitz, N., Nieman, D.H., Becker, H.E., van Amelsvoort, T.A.M.J., Dingemans, P.M., Linszen, D.H., de Haan, L., 2011. Semantic fluency deficits and reduced grey matter before transition to psychosis: a voxelwise correlational analysis. Psychiatry Res. 194, 1–6. doi:10.1016/j.pscychresns.2011.01.004

Mortamet, B., Bernstein, M.A., Jack, C.R., Gunter, J.L., Ward, C., Britson, P.J., Meuli, R., Thiran, J.-P., Krueger, G., Alzheimer's Disease Neuroimaging Initiative, 2009. Automatic quality assessment in structural brain magnetic resonance imaging. Magn. Reson. Med. 62, 365–372. doi:10.1002/mrm.21992

Pardoe, H.R., Kucharsky Hiess, R., Kuzniecky, R., 2016. Motion and morphometry in clinical and nonclinical populations. NeuroImage 135, 177–185. doi:10.1016/j.neuroimage.2016.05.005

Pfefferbaum, A., Mathalon, D.H., Sullivan, E.V., Rawles, J.M., Zipursky, R.B., Lim, K.O., 1994. A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. Arch. Neurol. 51, 874–887.

Pizarro, R.A., Cheng, X., Barnett, A., Lemaitre, H., Verchinski, B.A., Goldman, A.L., Xiao, E., Luo, Q., Berman, K.F., Callicott, J.H., Weinberger, D.R., Mattay, V.S., 2016. Automated Quality Assessment of Structural Magnetic Resonance Brain Images Based on a Supervised Machine Learning Algorithm. Front. Neuroinformatics 10, 52. doi:10.3389/fninf.2016.00052

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59, 2142–2154. doi:10.1016/j.neuroimage.2011.10.018

Power, J.D., Schlaggar, B.L., Petersen, S.E., 2015. Recent progress and outstanding issues in motion correction in resting state fMRI. NeuroImage 105, 536–551. doi:10.1016/j.neuroimage.2014.10.044

Reiss, A.L., Abrams, M.T., Singer, H.S., Ross, J.L., Denckla, M.B., 1996. Brain development, gender and IQ in children. A volumetric imaging study. Brain J. Neurol. 119 (Pt 5), 1763–1774.

Reuter, M., Tisdall, M.D., Qureshi, A., Buckner, R.L., van der Kouwe, A.J.W., Fischl, B., 2015. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. NeuroImage 107, 107–115. doi:10.1016/j.neuroimage.2014.12.006

Revelle, W., 2017. psych: Procedures for Psychological, Psychometric, and Personality Research.

Roalf, D.R., Quarmley, M., Elliott, M.A., Satterthwaite, T.D., Vandekar, S.N., Ruparel, K., Gennatas, E.D., Calkins, M.E., Moore, T.M., Hopson, R., Prabhakaran, K., Jackson, C.T., Verma, R., Hakonarson, H., Gur, R.C., Gur, R.E., 2016. The impact of quality assurance assessment on diffusion tensor imaging outcomes in a large-scale population-based cohort. NeuroImage 125, 903–919. doi:10.1016/j.neuroimage.2015.10.068

Roalf, D.R., Vandekar, S.N., Almasy, L., Ruparel, K., Satterthwaite, T.D., Elliott, M.A., Podell, J., Gallagher, S., Jackson, C.T., Prasad, K., Wood, J., Pogue-Geile, M.F., Nimgaonkar, V.L., Gur, R.C., Gur, R.E., 2015. Heritability of Subcortical and Limbic Brain Volume and Shape in Multiplex-Multigenerational Families with Schizophrenia. Biol. Psychiatry 77, 137–146. doi:10.1016/j.biopsych.2014.05.009

Satterthwaite, T.D., Connolly, J.J., Ruparel, K., Calkins, M.E., Jackson, C., Elliott, M.A., Roalf, D.R., Hopson, R., Prabhakaran, K., Behr, M., Qiu, H., Mentch, F.D., Chiavacci, R., Sleiman, P.M.A., Gur, R.C., Hakonarson, H., Gur, R.E., 2016. The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. NeuroImage, Sharing the wealth: Brain Imaging Repositories in 2015 124, Part B, 1115–1119. doi:10.1016/j.neuroimage.2015.03.056

Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M.E., Hopson, R., Jackson, C., Keefe, J., Riley, M., Mentch, F.D., Sleiman, P., Verma, R., Davatzikos, C., Hakonarson, H., Gur, R.C., Gur, R.E., 2014. Neuroimaging of the Philadelphia neurodevelopmental cohort. NeuroImage 86, 544–553. doi:10.1016/j.neuroimage.2013.07.064

Satterthwaite, T.D., Wolf, D.H., Loughead, J., Ruparel, K., Elliott, M.A., Hakon, H., Gur, R.C., Gur, R.E., 2012. Impact of In-Scanner Head Motion on Multiple Measures of Functional Connectivity: Relevance for Studies of Neurodevelopment in Youth. NeuroImage 60, 623–632. doi:10.1016/j.neuroimage.2011.12.063

Savalia, N.K., Agres, P.F., Chan, M.Y., Feczko, E.J., Kennedy, K.M., Wig, G.S., 2017. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. Hum. Brain Mapp. 38, 472–492. doi:10.1002/hbm.23397

Shaw, P., Kabani, N.J., Lerch, J.P., Eckstrand, K., Lenroot, R., Gogtay, N., Greenstein, D., Clasen, L., Evans, A., Rapoport, J.L., Giedd, J.N., Wise, S.P., 2008. Neurodevelopmental trajectories of the human cerebral cortex. J. Neurosci. Off. J. Soc. Neurosci. 28, 3586–3594. doi:10.1523/JNEUROSCI.5309-07.2008

Shehzad, Z., Giavasis, S., Li, Q., Benhajali, Y., Yan, C., Yang, Z., Milham, M., Bellec, P., Craddock, C., n.d. The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of MRI data. Front. Neurosci. doi:10.3389/conf.fnins.2015.91.00047

Sikka, S., Cheung, B., Khanuja, R., Ghosh, S., Yan, C., Li, Q., Vogelstein, J., Burns, R., Colcombe, S., Craddock, C., Mennes, M., Kelly, C., Dimartino, A., Castellanos, F., Milham, M., n.d. Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC). Front. Neuroinformatics. doi:10.3389/conf.fninf.2014.08.00117

Smith, T.B., Nayak, K.S., 2010. MRI artifacts and correction strategies. Imaging Med. 2, 445–457. doi:10.2217/iim.10.33

Sobel, M.E., 1982. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociol. Methodol. 13, 290–312. doi:10.2307/270723
Sowell, E.R., Peterson, B.S., Thompson, P.M., Welcome, S.E., Henkenius, A.L., Toga, A.W., 2003. Mapping cortical change across the human life span. Nat. Neurosci. 6, 309–315. doi:10.1038/nn1008
Sowell, E.R., Thompson, P.M., Leonard, C.M., Welcome, S.E., Kan, E., Toga, A.W., 2004. Longitudinal mapping of cortical thickness and brain growth in normal children. J. Neurosci. Off. J. Soc. Neurosci. 24, 8223–8231. doi:10.1523/JNEUROSCI.1798-04.2004
Sowell, E.R., Thompson, P.M., Tessner, K.D., Toga, A.W., 2001. Mapping Continued Brain Growth and Gray Matter Density Reduction in Dorsal Frontal Cortex: Inverse Relationships during Postadolescent Brain Maturation. J. Neurosci. 21, 8819–8829.
Styner, M.A., Charles, H.C., Park, J., Gerig, G., 2002. Multisite validation of image analysis methods: assessing intra- and intersite variability. pp. 278–286. doi:10.1117/12.467167
Tamnes, C.K., Ostby, Y., Fjell, A.M., Westlye, L.T., Due-Tønnessen, P., Walhovd, K.B., 2010. Brain maturation in adolescence and young adulthood: regional age-related changes in cortical thickness and white matter volume and microstructure. Cereb. Cortex N. Y. N 1991 20, 534–548. doi:10.1093/cercor/bhp118
Tisdall, M.D., Hess, A.T., Reuter, M., Meintjes, E.M., Fischl, B., van der Kouwe, A.J.W., 2012. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. Magn. Reson. Med. 68, 389–399. doi:10.1002/mrm.23228
Tisdall, M.D., Reuter, M., Qureshi, A., Buckner, R.L., Fischl, B., van der Kouwe, A.J.W., 2016. Prospective motion correction with volumetric navigators (vNavs) reduces the bias and variance in brain morphometry induced by subject motion. NeuroImage 127, 11–22. doi:10.1016/j.neuroimage.2015.11.054
Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: Improved N3 Bias Correction. IEEE Trans. Med. Imaging 29, 1310–1320. doi:10.1109/TMI.2010.2046908
Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., van Strien, N., Stone, J.R., Gee, J.C., Avants, B.B., 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. NeuroImage 99, 166–179. doi:10.1016/j.neuroimage.2014.05.044
Van Dijk, K.R.A., Sabuncu, M.R., Buckner, R.L., 2012. The influence of head motion on intrinsic functional connectivity MRI. NeuroImage 59, 431–438. doi:10.1016/j.neuroimage.2011.07.044
Vincent, J.L., Kahn, I., Snyder, A.Z., Raichle, M.E., Buckner, R.L., 2008. Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. J. Neurophysiol. 100, 3328–3342. doi:10.1152/jn.90355.2008
Wang, B., 2015. bda: Density Estimation for Grouped Data.
Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-Atlas Segmentation with Joint Label Fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35, 611–623. doi:10.1109/TPAMI.2012.143
Woodard, J.P., Carley-Spencer, M.P., 2006. No-Reference image quality metrics for structural MRI. Neuroinformatics 4, 243–262. doi:10.1385/NI:4:3:243
Yan, C.-G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R.C., Di Martino, A., Li, Q., Zuo, X.-N., Castellanos, F.X., Milham, M.P., 2013. A comprehensive assessment of regional

variation in the impact of head micromovements on functional connectomics. NeuroImage 76, 183–201. doi:10.1016/j.neuroimage.2013.03.004
Yendiki, A., Koldewyn, K., Kakunoori, S., Kanwisher, N., Fischl, B., 2014. Spurious group differences due to head motion in a diffusion MRI study. NeuroImage 88, 79–90. doi:10.1016/j.neuroimage.2013.11.027
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57. doi:10.1109/42.906424

TABLES

Table 1: Demographic information of the training and validation datasets.

Study	N	% Female	Age Mean	Age SD
Training	1067	0.51	14.90	3.70
Validation	534	0.55	15.10	3.68

Table 2: Quantitative quality metrics and associated citations.

Quantitative Metric	Abbreviation	Citation
signal-to-noise ratio	SNR	Magnotta and Friedman, 2006
contrast-to-noise ratio	CNR	Magnotta and Friedman, 2006
Foreground-to-background energy ratio	FBER	-
percent of artifactual voxels	Qi1	Mortamet et al., 2009
smoothness of voxels	FWHM	Friedman et al., 2006
entropy focus criterion	EFC	Atkinson et al., 1997

Table 3: Inclusion model results.

	Specificity	Sensitivity
Training	0.94	0.80
Validation	1	0.78

Table 4: Quantification model results

	Specificity	Sensitivity
Training	0.74	0.86
Validation	0.69	0.82

FIGURE LEGENDS

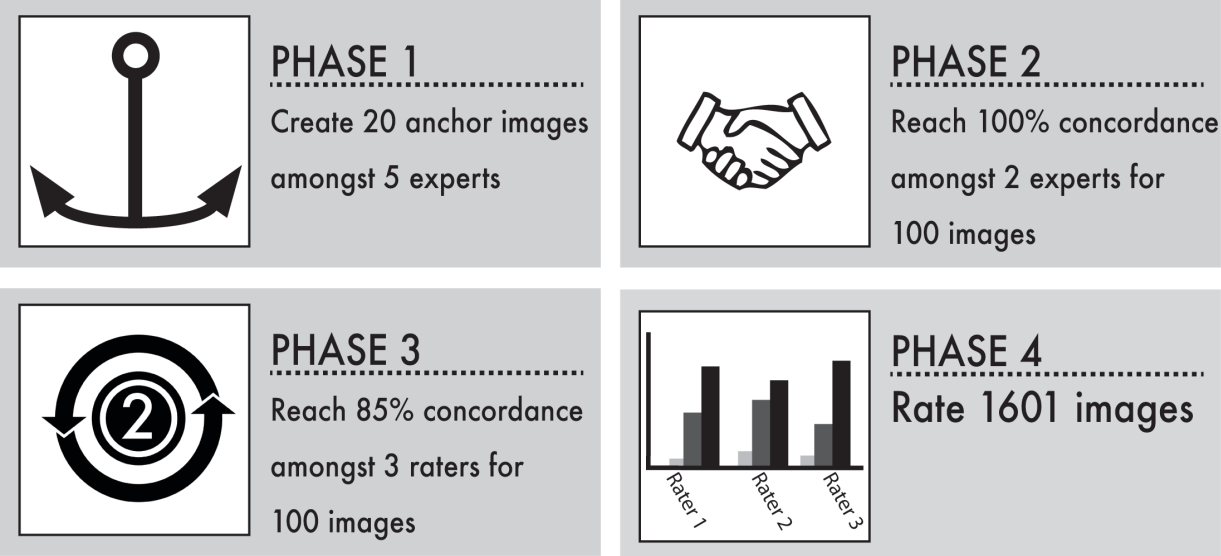

Figure 1: **Training protocol for manual raters.** There were 4 phases of training. ***Phase 1***: 5 neuroimaging experts reviewed 20 PNC images selected to have various types of artifact. These images were used to establish rating anchors, which were then used for Phase 2. ***Phase 2***: Two experts (TDS & DRR) rated 100 PNC images. 100% concordance was achieved through consensus. ***Phase 3***: Three new raters were trained on the 100 images used in Phase 2, until the raters achieved 85% concordance after two rounds. ***Phase 4***: All 3 trained raters manually rated 1,601 PNC images.

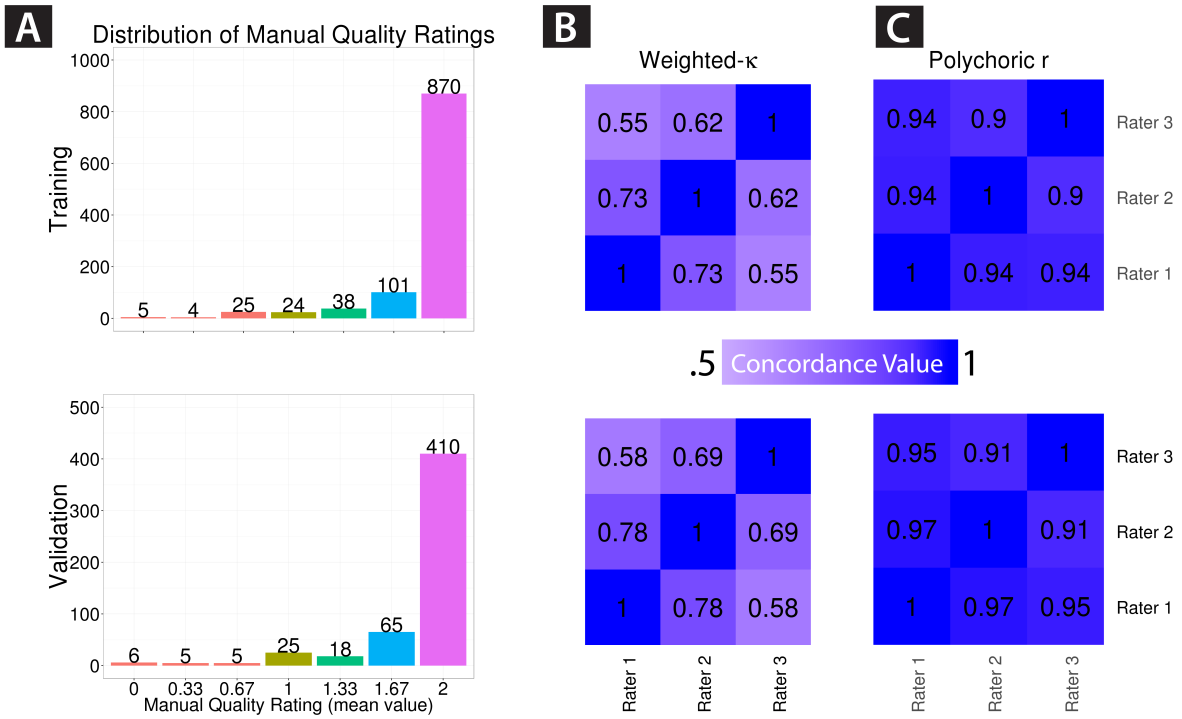

Figure 2: Results of manual ratings. ***A*****:** Frequency of average manual rating for the training (n = 1067) and validation datasets (n = 534). ***B*****:** The pairwise weighted-κ between each rater in the training and validation sets was moderate. ***C*****:** The pairwise polychoric correlation for each rater in both the training and validation sets was high.

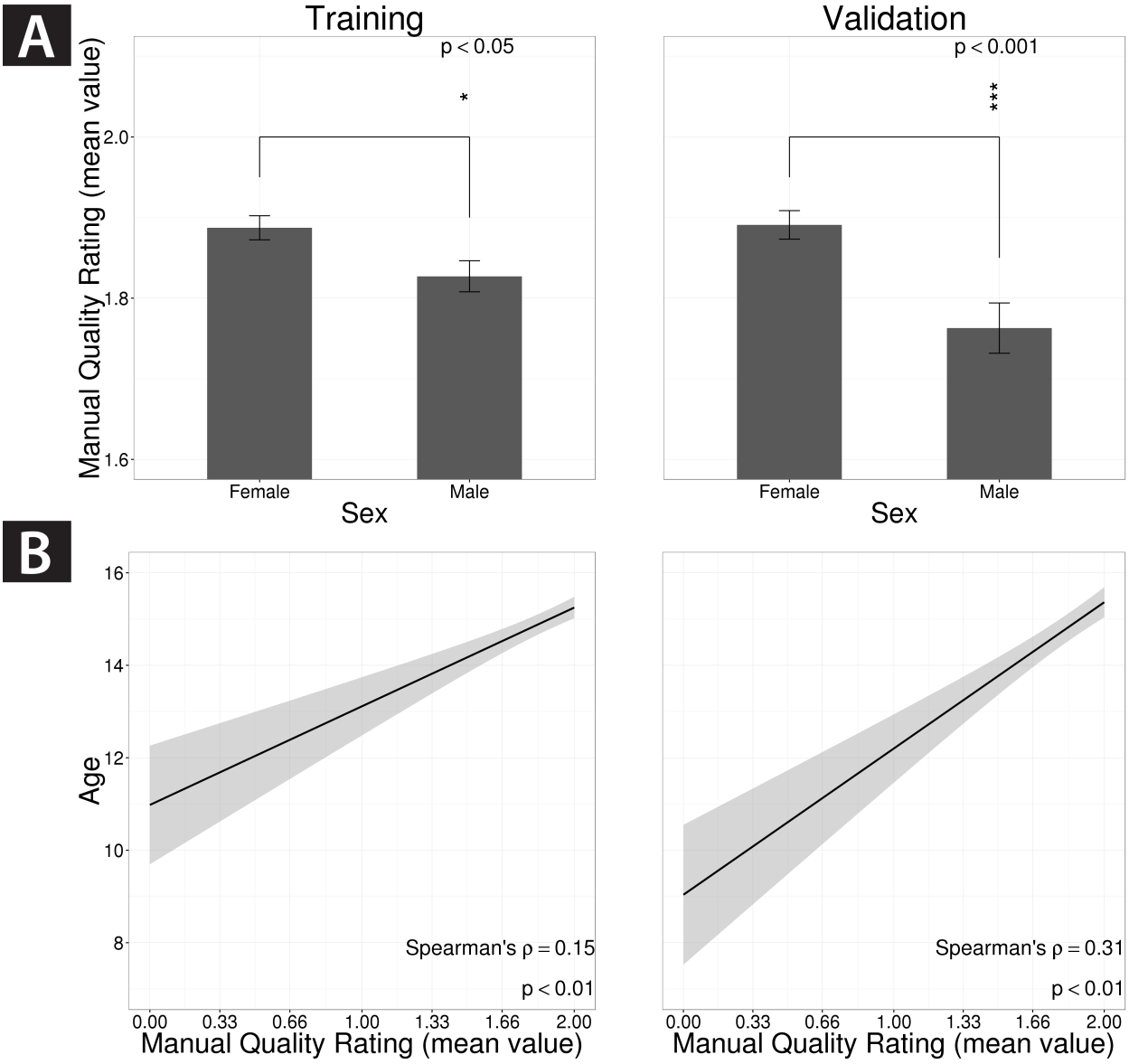

Figure 3: Manual rating varies by sex and is associated with participant age. ***A*:** Females have higher image quality than males in both the training and validation dataset. Bars represent the mean quality rating (+/- standard error of the mean); significance values are from Wilcoxon signed rank tests. ***B*:** Image quality improves with age in both training and validation datasets. Dark line represents a linear fit; shaded envelope represents 95% confidence intervals; significance values are from Spearman's correlations.

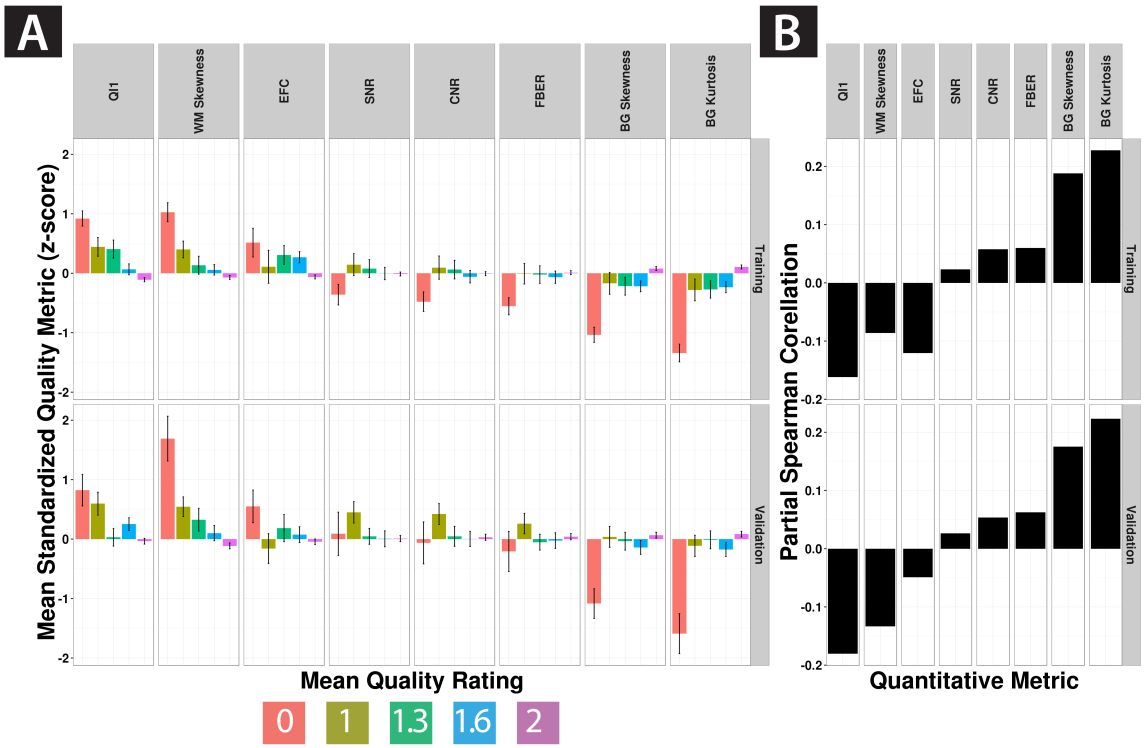

Figure 4: Quantitative metrics of image quality show heterogeneous alignment with manual ratings. ***A*:** The standardized mean (+/- S.E.M.) for each T1 derived QA metric is displayed by average manual rating. Several metrics including background (BG) kurtosis, BG skewness, and white matter skewness scaled with average manual quality rating. ***B*:** Partial Spearman correlation coefficients between average manual quality rating and the T1 derived quantitative metrics. Covariates included sex, age, and age squared.

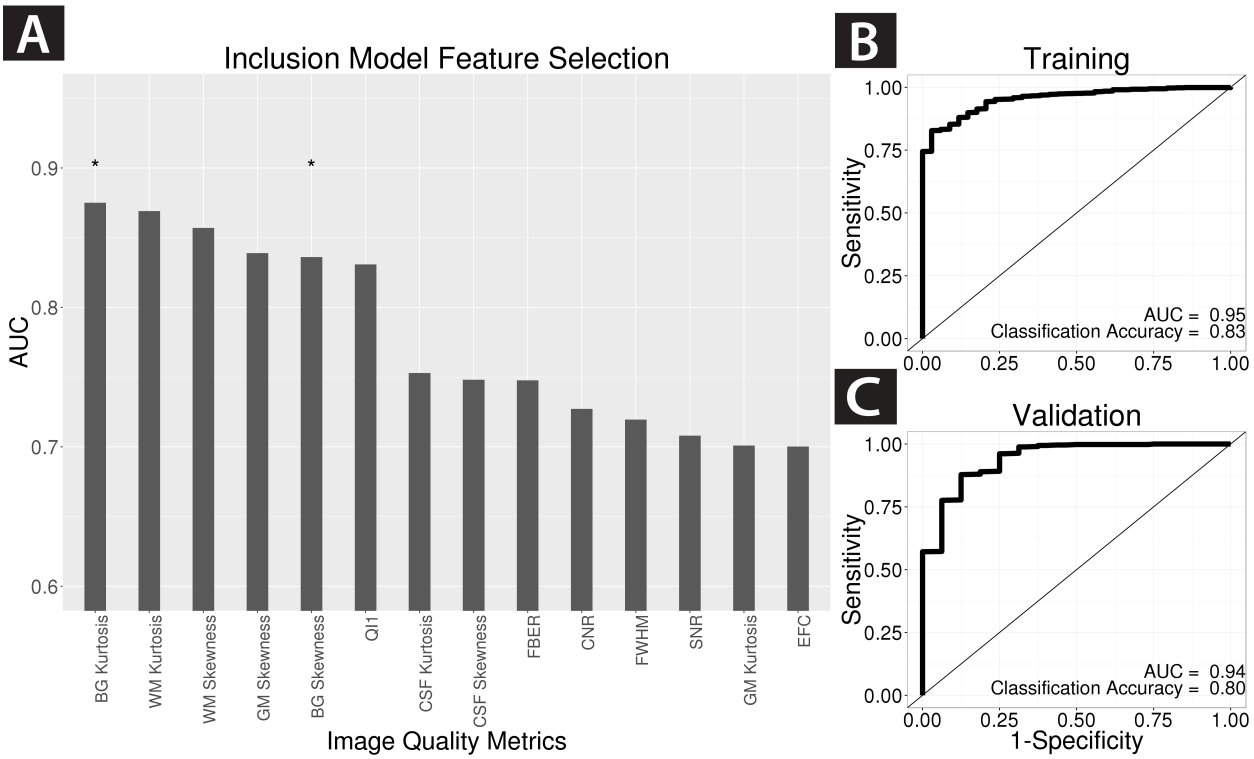

Figure 5: Inclusion model. The inclusion model aimed to differentiate usable and unusable images (manual rating of '0') using quantitative measures of image quality. ***A***: Variable selection for inclusion model. Mass-univariate AUCs are displayed for each quantitative metric in the training sample. The final two variables included in the model are marked by a star, these variables included background kurtosis and background skewness. Variables are ordered in decreasing AUC values for the initial mass-univariate analysis. ***B***: Inclusion model performance in training sample. The inclusion model displayed very high performance, with an AUC of 0.95 and a classification accuracy of 83%. ***C***: Inclusion model performance in validation sample. Variables and criterion from training sample were applied to an independent validation dataset. Model performance remained high in this new dataset, with an AUC of 0.94 and a classification accuracy of 80%.

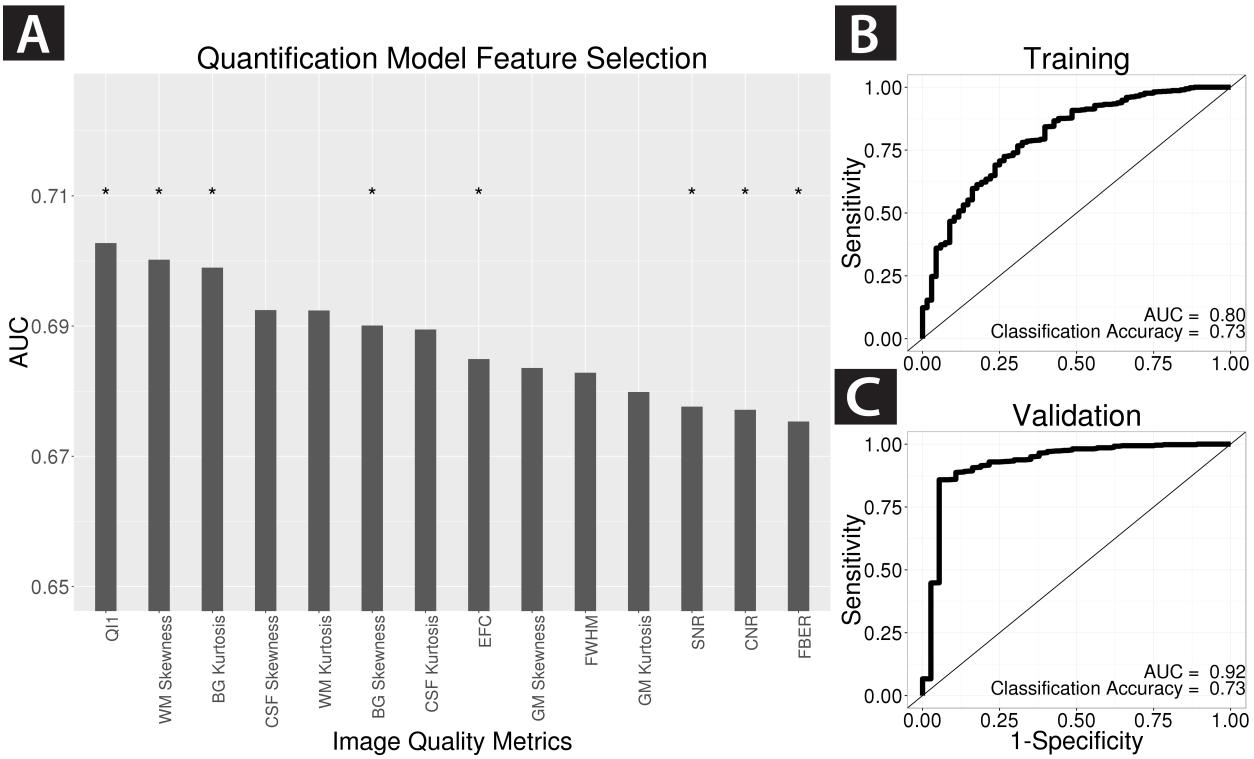

Figure 6: Quantification model. The quantification model aimed to identify images with some artifact (manual rating of '1') from images without artifact (manual rating of '2'). ***A***: Variable selection for quantification model. Mass-univariate AUCs are displayed for each quantitative metric in the training sample; variables are ordered in decreasing AUC values in the initial mass-univariate analysis. The final eight variables included in the model are marked by a star, these variables are Qi1, white matter skewness, background kurtosis, background skewness, energy focus criterion, signal-to-noise ratio, contrast-to-noise ratio, and foreground-to-background energy. ***B***: Quantification model performance in training sample. The inclusion model displayed very good performance, with an AUC of 0.80 and a classification accuracy of 73%. ***C***: Quantification model performance in validation sample. Variables and criterion from training sample were applied to an independent validation dataset. Model performance was excellent in this new dataset, with an AUC of 0.92 and a classification accuracy of 73%.

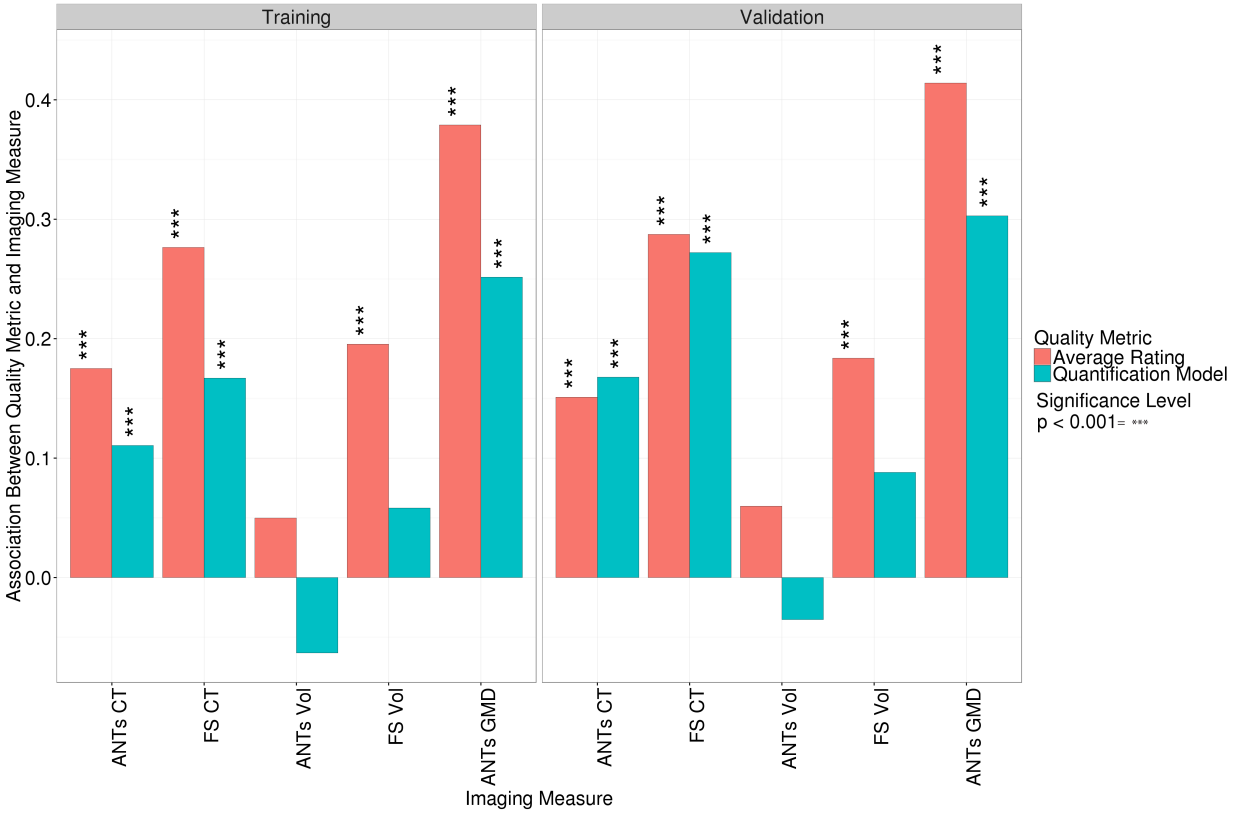

Figure 7: T1-weighted image quality is related to global measures of brain structure. Partial Spearman correlations are displayed between quality metrics and measures including mean cortical thickness (CT), mean cortical volume (Vol), and mean gray matter density (GMD). Thickness and volume were calculated using both ANTs and FreeSurfer. Covariates for these correlations included age, the quadratic effect of age, and sex. In general, quantitative measures of image quality aligned well with the average manual rating. Higher data quality was associated with greater mean cortical thickness; FreeSurfer had a stronger association with data quality than ANTs. Gray matter volume was least impacted by data quality, although the impact was again more pronounced for FreeSurfer. In contrast, gray matter density showed a strong relationship with data quality, with higher quality data being associated with greater density. Multiple comparisons were corrected for using FDR.

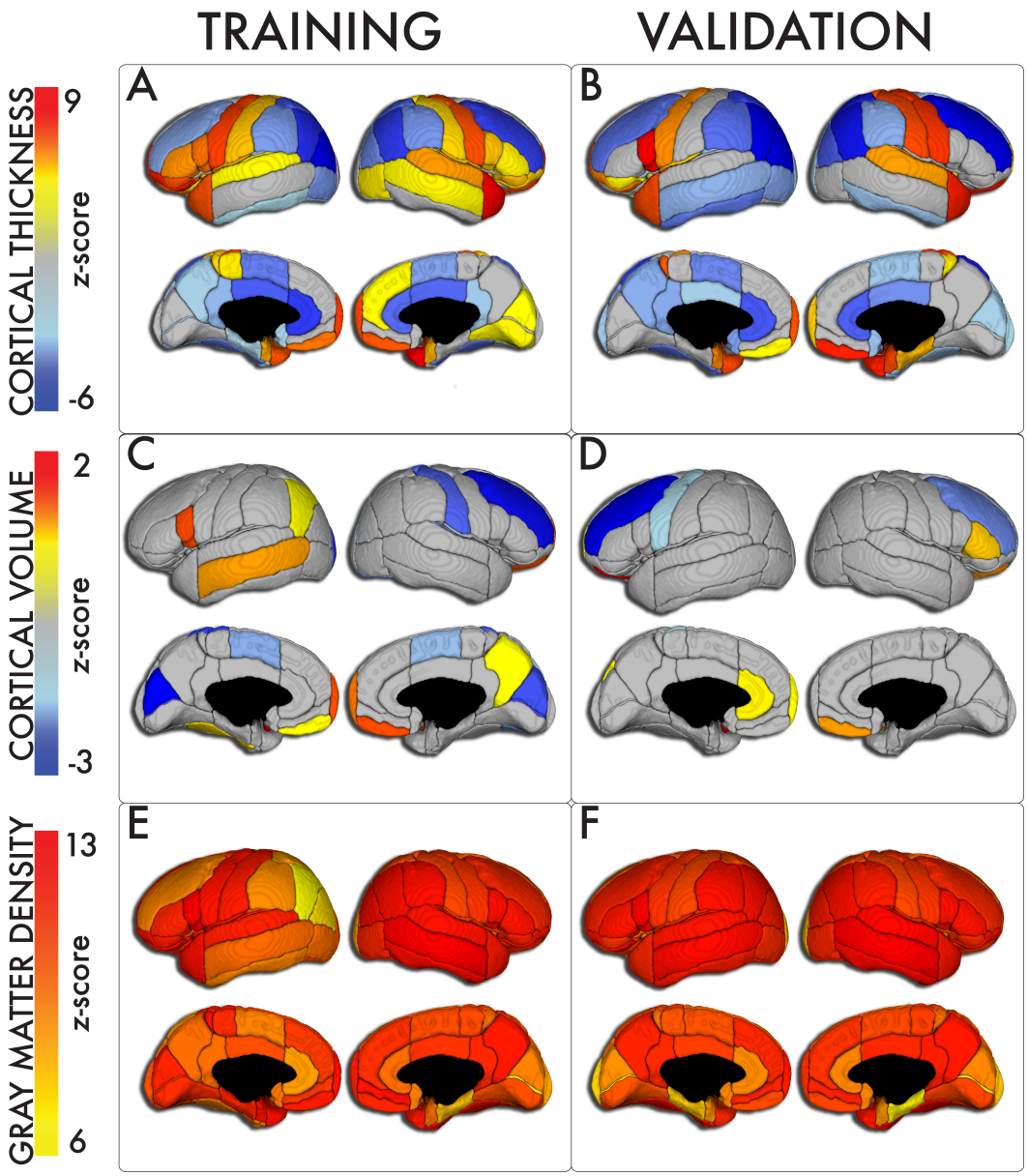

Figure 8: Measures of brain structure have divergent relationships to T1 quality. ***A-B:*** Cortical thickness has a regionally heterogeneous relationship with data quality. Higher data quality was associated with thicker cortex over much of the brain, but was conversely associated with thinner cortex in the dorsolateral prefrontal cortex, superior parietal cortex, and cingulate cortex. Image displays *z*-scores from a linear regression that modeled the cortical thickness as predicted by the $\hat{y}$ from the quantification model; covariates included age, age squared, and sex. All results corrected for multiple comparisons using the False Discovery Rate ($Q < 0.05$). ***C-D*:** Cortical volume is less strongly associated with quantitative measures of data quality. ***E-F*:** Gray matter density displayed the strongest association with data quality. Higher data quality was associated with greater gray matter density in both the training and validation datasets, where all regions survived FDR correction.

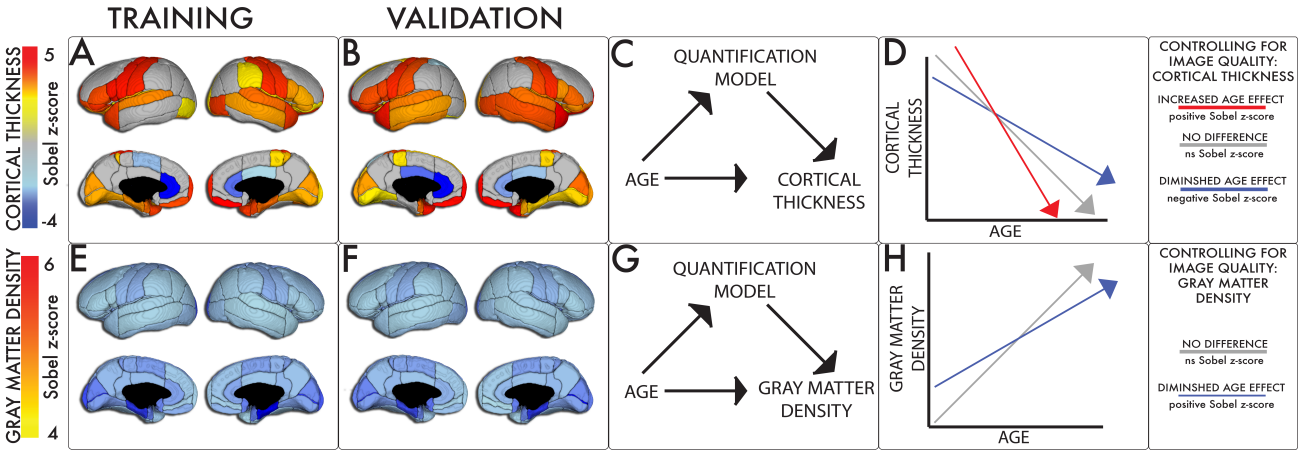

Figure 9: Data quality significantly mediates the apparent development of brain structure with age. ***A-B, E-F:*** A mediation analysis was performed for each cortical region. Sobel *z*-scores were calculated from the output of each mediation analysis; FDR correction was performed and only significant mediation effects are displayed. Data quality significantly mediated the relationship between age and cortical thickness (***C***) in a bidirectional, regionally heterogeneous manner (see schematic in panel ***D***). In general, data quality masked apparent associations with age. After controlling for data quality, the apparent age effect was increased in many regions (regions in warm colors in panels ***A*** and ***B***). However, in a subset of regions including the anterior cingulate, data quality inflated apparent associations with age, and controlling for data quality resulted in less significant age associations (cool colors in panels ***A*** and ***B***). In contrast, for gray matter density, the effects were global, with every region displaying a significant mediation effect, and quite consistent (***E*** and ***F***). Throughout the brain, data quality inflated apparent associations with age, and controlling for data quality resulted in a diminished association between age and gray matter density (***H***). All analyses controlled for multiple comparisons using FDR ($Q < 0.05$).

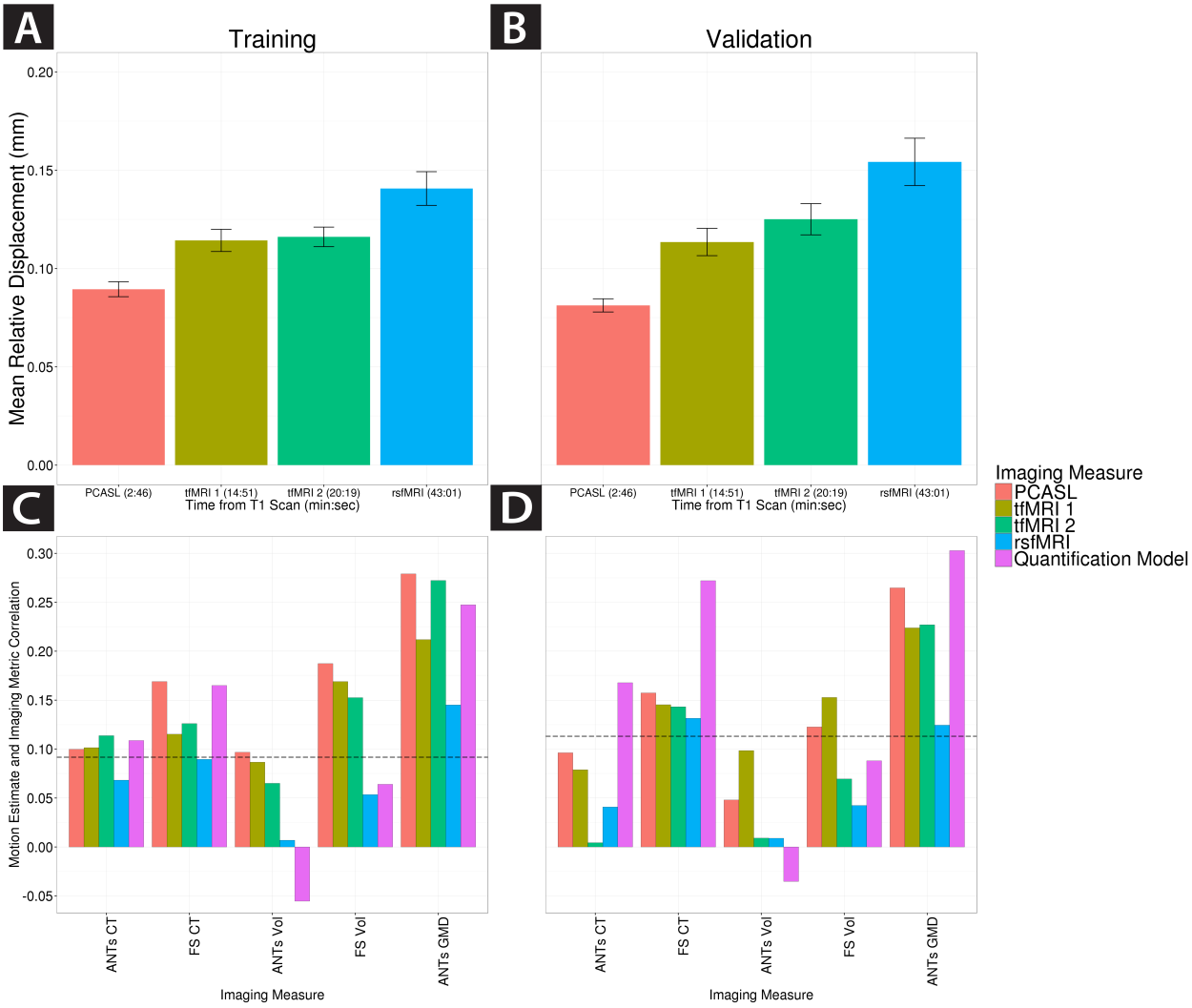

Figure 10: Association between motion during functional sequences and global measures of brain structure. *A-B*: Motion during functional sequences increases as scanning session progresses. Bars display mean frame displacement (+/- SEM) of motion during each function scan. ***C-D*:** Motion during functional sequences shows a similar relationship to global measures of brain structure as quantitative measures of data quality. Functional sequences that are acquired closer in time to the structural image tend to show a stronger association with each measure. Bars display the partial correlation between global measures of brain structure (as in Figure 7) and motion during each functional sequence. The partial correlation between motion and imaging metrics is displayed as the absolute value in order to align polarity of effect with the quantification model. The correlation value required to be significant at an FDR corrected $p < .05$ is plotted as the horizontal dashed line.