Received: 1 March 2017 | Revised: 22 November 2017 | Accepted: 29 November 2017

DOI: 10.1002/hbm.23911

RESEARCH ARTICLE

WILEY

Automated quality assessment of structural magnetic resonance images in children: Comparison with visual inspection and surface-based reconstruction

Tonya White[1,2] | Philip R. Jansen[1,2,3] | Ryan L. Muetzel[1,3] | Gustavo Sudre[4] | Hanan El Marroun[1,3,5] | Henning Tiemeier[1,5,6] | Anqi Qiu[7,8] | Philip Shaw[4] | Andrew M. Michael[9] | Frank C. Verhulst[1,10]

[1]Department of Child and Adolescent Psychiatry, Erasmus University Medical Centre, Rotterdam, Netherlands

[2]Department of Radiology, Erasmus University Medical Centre, Rotterdam, Netherlands

[3]The Generation R Study Group, Erasmus University Medical Centre, Rotterdam, Netherlands

[4]The Neurobehavioral Clinical Research Section, Social and Behavioral Research Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland

[5]Department of Pediatrics, Erasmus University Medical Centre, Rotterdam, Netherlands

[6]Department of Epidemiology, Erasmus University Medical Centre, Rotterdam, Netherlands

[7]Department of Biomedical Engineering and Clinical Imaging Research Center, National University of Singapore, Singapore, Singapore

[8]Singapore Institute for Clinical Sciences, Singapore, Singapore

[9]Autism and Developmental Medicine Institute, Geisinger Health System, Lewisburg, Pennsylvania 17837

[10]Department of Clinical Medicine at the Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Correspondence
Tonya White, MD, PhD, Department of Child and Adolescent Psychiatry, Erasmus MC – Sophia Children's Hospital, Dr. Molewaterplein 60/kamer Kp-2869, 3000 CB Rotterdam, Netherlands.
Email: t.white@erasmusmc.nl

Funding information
Netherlands Organization for Health Research and Development (ZonMw) TOP project number: 91211021; European Community's 7th Framework Programme (FP7/2008–2013), Grant/Award Number: 212652; Stichting Sophia Kinderziekenhuis Fonds; Erasmus Medical Center, Rotterdam; Erasmus University Rotterdam; Netherlands Organization for Scientific Research (NWO); Ministry of Health, Welfare and Sport; Ministry of Youth and Families

Abstract

Motion-related artifacts are one of the major challenges associated with pediatric neuroimaging. Recent studies have shown a relationship between visual quality ratings of T_1 images and cortical reconstruction measures. Automated algorithms offer more precision in quantifying movement-related artifacts compared to visual inspection. Thus, the goal of this study was to test three different automated quality assessment algorithms for structural MRI scans. The three algorithms included a Fourier-, integral-, and a gradient-based approach which were run on raw T_1-weighted imaging data collected from four different scanners. The four cohorts included a total of 6,662 MRI scans from two waves of the Generation R Study, the NIH NHGRI Study, and the GUSTO Study. Using receiver operating characteristics with visually inspected quality ratings of the T_1 images, the area under the curve (AUC) for the gradient algorithm, which performed better than either the integral or Fourier approaches, was 0.95, 0.88, and 0.82 for the Generation R, NHGRI, and GUSTO studies, respectively. For scans of poor initial quality, repeating the scan often resulted in a better quality second image. Finally, we found that even minor differences in automated quality measurements were associated with FreeSurfer derived measures of cortical thickness and surface area, even in scans that were rated as good quality. Our findings suggest that the inclusion of automated quality assessment measures can augment visual inspection and may find use as a covariate in analyses or to identify thresholds to exclude poor quality data.

KEYWORDS

artifacts, FreeSurfer, head movement, pediatric neuroimaging, pediatric population neuroscience, population neuroscience, quality assurance

1 | INTRODUCTION

Imaging artifacts remain a common challenge for neuroimaging studies, especially in children and specific clinical populations. While artifacts in some sequences have the advantage of becoming a contrast medium in other sequences (i.e., diffusion or flow), other artifacts, such as motion artifacts, remain problematic in image analyses. In fact, motion related artifacts are one of the major challenges associated with imaging pediatric populations (Blumenthal, Zijdenbos, Molloy, & Giedd, 2002, Backhausen et al., 2016) and has received considerable attention due to the influence of motion on connectivity-based analyses of resting-state fMRI (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012, Satterthwaite et al., 2012, Van Dijk, Sabuncu, & Buckner, 2012). However, motion-related issues also impact structural MRI with evidence that movement related artifacts can influence measures such as volumes of cortical and deeper structures (Blumenthal et al., 2002, Alexander-Bloch et al., 2016) and cortical thickness and surface area (Reuter et al., 2015, Backhausen et al., 2016, Ducharme et al., 2016). Thus, metrics that can accurately quantify movement and other artifacts are important to not only select images that should be excluded, but also to potentially statistically correct for minor movements that can influence the morphologic variables. This is supported by a recent study finding that movement related artifacts affect cortical thickness, even after removal of scans that failed a stringent visual quality control procedure (Reuter et al., 2015).

Quality control is especially important as the neuroimaging field moves to increasingly larger sample sizes, with an exponentially increasing number of scans that require ratings. Structural scans typically undergo multiple visual assessment steps, such as an initial inspection for incidental findings, inspection to determine raw T_1-weighted image quality, and following image processing with tools such as FreeSurfer to assure optimum segmentation and surface reconstruction (El Marroun et al., 2014, Backhausen et al., 2016). Multiple levels of visual inspection is not only time consuming, but errors can occur due to rater drift and difference in raters, resulting in a decrease of intra- and inter-rater reliability, respectively. Additional errors can occur in multi-site studies resulting from differences in raters and rating algorithms across multiple sites, since sites have an “institutional history” associated with how Q/A is performed. While visual inspection should always be performed, the development of automated algorithms to quantify image quality can be complementary to visual inspection. Automated algorithms have the advantage of not being prone to rater drift or performance differences between raters and also can provide more precise quality measurements. For example, Gardner et al. (1995) used images that were manipulated in such a way as to systematically alter cortical thickness to test the sensitivity of human raters. They found that visual raters were able to accurately detect changes of cortical thickness of about 40%, thus automated approaches should be able to detect more subtle differences compared to visual-inspection.

There are several different approaches in the literature to automatically derive quality assessment metrics from structural MRI scans (Atkinson, Hill, Stoyle, Summers, & Keevil, 1997, Mortamet et al., 2009, Pizarro et al., 2016), and available software to generate quality assessment metrics (http://preprocessed-connectomes-project.org/quality-assessment-protocol/index.html). Mortamet et al. (2009) measured voxel intensities in the background noise with the hypothesis that artifacts cause a right-skew in the distribution of voxel intensities. This approach was tested on a group of 188 elderly subjects from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008) with good results. Pizarro et al. (2016) challenged this approach stating that one metric alone is not sufficient to capture the number of artifacts present in structural neuroimaging data. They presented findings from a machine learning approach where different features were extracted from structural imaging data. They reported a sensitivity and specificity of their support vector machine (SVM) approach of 70.1% and 88.2%, respectively. However, it is unclear to what extent different MR platforms affected the accuracy.

Given the importance to obtain automated metrics for the quality of structural MRIs, the goal of this study was to develop and compare three different approaches (two novel and one variation of the approach by Mortamet et al., 2009) to measure the quality of structural MRI scans using four large cohorts that, while focusing on pediatric populations, cover the lifespan. Two of the metrics tested were based on the properties that voxels collected in k-space will contribute to the entire field-of-view following a Fourier transform from k-space to image space. Thus, movement of a subject during scanning results in “waves,” or banding, on the image seen in the spatial domain. The first two algorithms evaluate the noise characteristics outside the head, in the air, starting several voxels outside the head and extending laterally. The first approach utilizes a Fourier transform to capture the frequency characteristics of the noise ripples away from the edge of the head (ringing). The second metric calculates the integral of the voxel intensities characterized as vectors radiating away from the head, an approach similar to the method used by Mortamet et al. (2009). The third algorithm uses properties of the line spread function along the edge of the head, since movement is most prominent at the head/air interface (Barish and Jara, 1999). These three automated approaches were compared to systematic visual inspections in each of the four cohorts, of which intra- and inter-rater reliabilities of the visual inspections are reported.

Finally, we tested whether the automated quality assessment algorithm could predict the visually inspected quality of postprocessing reconstructions using FreeSurfer. Based on recent studies showing a relationship between automated derived measures cortical thickness and quality based on visual inspection, we also tested for the relationship between FreeSurfer-derived measures of cortical thickness and measures from the automated quality assessment. We hypothesized that automated algorithms will provide invaluable complementary information for visual inspection, with higher resolution metrics of quality that will provide a more accurate and reproducible threshold to exclude poor quality images from structural analyses.

2 | METHODS

2.1 | Subjects

2.1.1 | Generation R cohort

Images included in this study were acquired from children who were participants of the first and second neuroimaging waves of the

Generation R Study. The Generation R Study is a large, ethnically diverse epidemiological study of child development (Jaddoe et al., 2006, Tiemeier et al., 2012). The first neuroimaging wave of the Generation R Study began in September 2009 until July 2013 and a total of 1070 six- to nine-year-old children were scanned (White et al., 2013). The second neuroimaging wave started in April 2013 with 4,087 nine- to eleven-year-old children scanned (White et al., 2017). Prior to recruitment during each phase of the study informed consent was obtained and each neuroimaging wave was approved by the Medical Ethical Committee (METC). Exclusion criteria included contraindications for the MRI procedure (i.e., pacemaker, ferrous metal implants), claustrophobia, having a significant motor or sensory disorder, moderate to severe head trauma with loss of consciousness, and neurological disorders (including seizure disorder, neuromotor disorder, or a history of brain tumors). A total of 3,959 of the children have both parental consent and a complete T_1-weighted image.

2.1.2 | NHGRI cohort

Participants lacked any psychiatric diagnoses, as determined by DSM-5 based, clinician-administered interviews: for adults: the *Structured Clinical Interview for* DSM-IV-TR *Axis I Disorder Re-search Version, Patient Edition* and the *Conners' Adult ADHD Diagnostic Interview for* DSM-IV, for children, the parental Diagnostic Interview for Children and Adolescents-IV (DICA). Contraindications included major neurological disorders, substance dependence, and contraindications to MRI scanning. Adult participants provided written consent; children (under 18 years) gave written assent and their parents provided written consent for their child. All study procedures were approved by the Institutional Review Board of the National Human Genome Research Institute.

2.1.3 | GUSTO cohort

The GUSTO cohort recruited pregnant Singapore citizens or Permanent Residents of Chinese, Malay or Indian ethnic backgrounds from two major birthing hospitals in Singapore at the first antenatal visit. The cohort description is detailed in Soh et al. (2012). Children were recruited during the 4-year home visit of the GUSTO study and underwent MRI scans at ~4.5 years of age (± 1 month). The GUSTO study was approved by the National Healthcare Group Domain Specific Review Board (NHG DSRB) and the Sing Health Centralized Institutional Review Board (CIRB). Written informed consent was obtained from mothers prior to inclusion into the study.

2.2 | Magnetic resonance imaging

2.2.1 | Generation R study

Prior to the actual MRI, the children were familiarized with the MRI procedure during a mock scanner session. During the MRI scan, care was taken so that the children rested comfortably in the scanner and soft cushions were used to assist with head immobilization. The children were able to watch a film of their choice during the acquisition and the film was projected onto a screen at the front of the scanner and the children watched though forward-directed mirrors. To motivate children to lie still in the scanner, we showed them an image of a brain with a lot of movement artifacts and no movement artifacts. The MR images for the first and second waves were collected on two different GE 3-Tesla scanners. The first wave was collected on a GE 750 Discovery clinical MR system using an 8-channel head coil and a T_1-weighted inversion recovery fast spoiled gradient recalled (IR-FSPGR) sequence. The following sequence parameters were used: TR = 10.3 ms, TE = 4.2 ms, TI = 350 ms, NEX = 1, flip angle = 16°, readout bandwidth = 20.8 kHz, matrix 256 × 256, imaging acceleration factor of 2, and an isotropic resolution of 0.9 × 0.9 × 0.9 mm^3. The total scan time for the T_1 was 5 min 40 s. The total sample of wave 1 was 1,070 six- to nine-year-old children.

MR images for the second neuroimaging wave were acquired on a research-dedicated GE 750w Discovery wide-bore MRI system (Milwaukee, MI, USA) using an 8-channel head coil. A high-resolution T_1-weighted sequences were obtained using a three-dimensional (3D) coronal inversion recovery fast spoiled gradient recalled (IR-FSPGR, BRAVO) sequence (T_R = 8.77 ms, T_E = 3.4 ms, T_I = 600 ms, flip angle = 10°, Field of view = 220 mm × 220 mm, number of slices = 230, voxel size = 1.0 mm^3, ARC acceleration = 2. A small subgroup of children (n = 21) had scans acquired at the beginning of the study using ASSET acceleration rather than ARC.

2.2.2 | NHGRI cohort

Participants were acclimatized to the scanning environment, rested comfortably in the scanner with the head immobilized. and could watch a film of their choice. A high-resolution (1.07 × 1.07 × 1.2 mm) T1 weighted volumetric structural image was obtained using a magnetization prepared rapid gradient echo sequence (with ASSET preparation; 124 slices, 1.2 mm slice thickness, 224 × 224 acquisition metric, flip angle = 6°, field of view = 24 cm^2) on a 3 T General Electric Signa scanner (USA) using an eight-channel head coil.

2.2.3 | GUSTO cohort

MRI scanning was performed using a 3 T Siemens Skyra system with a 32-channel head coil at KK Women's and Children's hospital. Children went through a MRI home training program prior to the MRI visit and on-site MRI training. Structural imaging involved a high-resolution T_1-weighted Magnetization Prepared Rapid Gradient Recalled Echo (MPRAGE; 192 slices, 1 mm thickness, in-plane resolution 1 mm, sagittal acquisition, field of view 192 × 192 mm, matrix = 192 × 192, repetition time = 2000 ms, echo time = 2.08 ms, inversion time = 877 ms, flip angle = 9°, scanning time = 3.5 min).

3 | IMAGE PROCESSING

Structural images from all three cohorts were processed using the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). Cortical and subcortical segmentation and surface reconstruction of the T_1-weighted images was performed using *recon all* from the Freesurfer Wave 1 was performed with FreeSurfer version 5.3 and wave 2 with FreeSurfer version 6.0. The technical details of these procedures have been described in detail in previous work (Dale, Fischl, & Sereno, 1999; Fischl, Sereno, & Dale, 1999a, Fischl, 2012). Briefly, this process included the removal of non-brain tissue (Segonne, Pacheco, & Fischl, 2004), automated Talairach transformation into standard space, intensity normalization (Fischl et al., 2004), tessellation of the gray/white matter boundary, automated topology correction (Segonne et al., 2007), and surface deformation (Fischl et al., 1999a). Once the cortical models were complete, the images underwent surface inflation (Fischl et al., 1999a), registration to a spherical atlas (Fischl, Sereno, Tootell, & Dale, 1999b), and the parcellation of the cerebral cortex into units based on gyral and sulcal structure (Desikan et al., 2006). Cortical thickness was calculated as the distance from the gray/white matter boundary to the gray matter/cerebral spinal fluid boundary at each vertex on the tessellated surface (Fischl and Dale, 2000). After running the standard processing steps of FreeSurfer, we calculated the mean cortical thicknesses of parcelated regions defining the frontal, temporal, parietal, and occipital lobes, bilaterally.

3.1 | Manual quality assessment of the T_1 images

3.1.1 | Generation R study

At the time of the MRI acquisition, T_1 images were evaluated for incidental findings and rated for image quality using a six-point Likert scale (Jansen, van der Lugt, & White, 2017). The quality assessment levels for the scans were: unusable, poor, fair, good, very good, and excellent. The visual inspection measures used to make this assessment included the sharpness of the gray matter and white matter interface on the cortex, the presence of ringing in the image, and whole brain coverage. If the initial T_1 scan was rated as unusable or poor by the technician or PhD student running the scanner, the T_1 sequence was repeated. Prior to repeating the scan, communication took place between the child and MR technician to make sure that the child was comfortable in the scanner and to remind the child to remain as still as possible.

3.1.2 | NHGRI cohort

All T1 images were visually inspected at the time of the scan (by PS). If the scan was felt to have more than minimal motion artefact, a second attempt was made and if motion persisted, the participant was offered a repeat scan at a later date in the evening (to increase the chance of scanning during natural sleep). The best quality image was then further rated as having no, mild, moderate or severe motion or other artifacts by two raters, using published guidelines (Blumenthal et al., 2002). Those judged by two raters to have no or minimal motion artefact proceeded to segmentation of cerebral cortical structures.

3.1.3 | GUSTO cohort

The T_1 images were rated for image quality at the time of scanning using a four-point Likert scale. The quality assessment levels for the scans were: unusable, large motion, minor motion, and no motion. The visual inspection measures used to make this assessment included the sharpness of the gray and white matter interface in the cortex, the presence of ringing in the image, and whole brain coverage. If the initial T_1 scan was rated as unusable or poor by the technician running the scanner, the T_1 sequence was repeated. Prior to repeating the scan, communication took place between the child and MR technician to make sure that the child was comfortable in the scanner. The usable scans were those rated as having either minor or no motion (Table 1).

3.2 | Manual quality assessment of the FreeSurfer images

3.2.1 | Generation R study

Wave 1—The 1,070 T_1-weighted images from the first neuroimaging wave underwent a thorough and systematic visual inspection to assess segmentation and surface quality. This was performed using a 7-point Likert scale with the following levels: not reconstructed, poor, fair, sufficient, good, very good, and excellent). Images rated as unusable or poor at the scan site, images that could not be processed by Freesurfer, and images with a poor segmentation quality were considered as failing the Q/A protocol. *Wave 2*—Of the 3,959 T_1-weighted images from the second wave 3,937 were reconstructed using FreeSurfer. All FreeSurfer reconstructions, including 2-D segmentations and 3-D morphometry were visually inspected using a 3-point Likert scale with the following levels: "Excellent to Very Good," "Good to Fair," and "Poor to Unusable."

3.2.2 | NHGRI cohort

Cerebral cortical reconstruction and cortical volumetric segmentation were performed with the FreeSurfer image analysis suite version 5.3.0 (http://surfer.nmr.mgh.harvard.edu/). Analyses were conducted on the National Institutes of Health High Performance Computer Cluster (Biowulf). These segmentations were inspected by two raters and scored following the ENIGMA guidelines. The 2-D segmentations were scored as "1" if no errors were detected; "2" if minor errors were noted; "3" if moderate errors were there; "4" if there were gross errors. If ratings differed by more than one point, the segmentations were reinspected and a consensus rating was reached.

3.2.3 | GUSTO cohort

Following Freesurfer guidelines, visual inspection of brain skullstripping, white matter, and pial surfaces was conducted. The manual correction, such as adding control points, based on FreeSurfer guideline was also performed.

3.3 | Automated T_1 quality assessment

A flow diagram of the algorithm used to automatically assess image quality is shown in Figure 1. The scripting and programming were performed in MATLAB (Version R2016a, Mathworks, Natick, MA); however, there is currently a beta Python version of the gradient approach available on Github (https://github.com/tjhwhite/auto_quality_assurace). The images were first converted from dicom to nifti using Dcm2Nii (http://lcni.uoregon.edu/downloads/mriconvert/mriconvert-and-mcverter). Next, a brain mask was created using FSL's brain extraction tool (BET2) (https://www.fmrib.ox.ac.uk/fsl) to identify the location and orientation of the brain in 3D space. The third step was to apply AFNI's 3dEdge3 (https://afni.nimh.nih.gov/afni/) to the raw T_1 image. AFNI's 3dEdge3 function is a three-dimensional edge detection

TABLE 1 Demographic and MRI ratings and automated quality assessment metrics for participants in the three neuroimaging cohorts

	Generation R Wave I	Generation R Wave II	NHGRI	GUSTO
Number of participants	1,070	3,940	442	252
Age (mean/SD) (years)	7.9/1.0	10.1/0.59	22.4 (14.7)	4.59 (.08)
Age range (years)	6.1–10.7	8.6–11.9	5.3–77.6	4.44–4.95
Sex (male/female)	572/498	1948/1992	252/190	118/134
Total number of scans	1,070	4339[a]	442	811[a]
Categorical scan quality	Excellent 227 Very good 349 Good 322 Fair 121 Poor 50 Unusable 1	Excellent 365 Good 2605 Fair 750 Poor 248 Unusable[a] 366	No motion 106 Mild motion 313 Moderate 19 Severe motion 4	Good 239 Minor motion 129 Large motion & unusable 443
Usable/unusable scans (T_1)	1019/51 (95.2%)	3,559/381 (89.3%)	419/23 (94.8%)	368/443 (44.8%)
Scans processed with FreeSurfer	1065	3923	442	811
Usable scans (FreeSurfer)	922	3234	345	252
Gradient automated Q/A metric				
- Total (mean/SD)	1103 (196)	1548 (152)	1346 (105)	114.4 (19.3)
- Usable scans	1118 (188)	1565 (129)	1356 (92)	123.0 (16.6)
- Unusable scans	796 (71)	1237 (197)	1164 (158)	107.4 (18.6)

[a]Includes repeat scans.

algorithm that returns an image with a clear outline of the interface between the outer border of the head and the air. However, we noted that in children with considerable movement, the edge detection failed to identify the edge of the head in a small number of regions. In these cases, the brain mask was used to find the missing borders of the head that were not detected via 3dEdge3. This step was performed automatically in the situation where the brain masked is reached before the edge when approaching the head from lateral to medial.

Within the 3D image field of view, the Euclidean distance was automatically calculated between the anterior, posterior, superior, right and left lateral borders of the head (defined by 3dEdge3) with each side of the field of view (FOV), respectively. This was performed to double-check that the whole head was captured within the image space (i.e., the child did not move part way out of the scanner's FOV). These Euclidean distance parameters were used to select a volume of the image outside of the head that was used to evaluate image quality. The region of interest (ROI) used for the analysis in the axial plane began 10 mm inferior from the top of the head, to 80 mm inferior to the top of the head. From a coronal section, the ROI plane started 50 mm back from the anterior portion of the brain to the plane that defined the posterior slice of the head (Figure 1; insert). Defining the plane as starting 50 mm posterior from the anterior portion of the head allowed for the removal of flow artifacts from eye movements.

Using the ROI defined above for the automated quality assessment (Q/A) analyses, linear one-dimensional vectors (one voxel extended laterally from the edge of the head sagitally and of length 100 voxels) were identified from T_1-weighted image. Combining these 1-D vectors, an array ($j \times k$) was constructed from the selected ROI (thus, along multiple slices of the image), where each j was an index for a different location on the edge of the head, and k was the length of the vector away from the head (set to $k = 100$). Thus, the array was a line of image intensities with element (j,1) marking the edge of the head at the point defined by 3dEdge3 (j could also be translated to an (x,y,z) coordinate system) to element (j,100) being 100 voxels lateral (right and left) from the edge of the head and extending outside the head. Using this array for each individual, three different approaches were used to quantify Q/A in the images from these j lines (100 voxels in length) from the edge of the head:

3.3.1 | Fourier

During scanning, data were collected in k-space and transformed to the spatial domain via a Fourier transform. Thus, head movement during scanning results in wave-like image artifacts in the spatial domain. Thus, our first approach was to assess the spatial frequency characteristics of voxels outside the head. For the Fourier transform algorithm, each k line of data underwent a Fourier transform starting 5 voxels outside the head. The 5 voxels provided a buffer from the rapid decline in voxel intensity that occurred along the air/head interface. Next, after removing the baseline component, the maximum to mean frequency was calculated by dividing the highest peak signal of the magnitude vector of the Fourier transform by the mean frequency across the spectrum. This ratio of max over mean quantified whether a characteristic frequency pattern dominates above a white noise pattern.

3.3.2 | Integral

The second approach took the integral of the noise outside of the brain along each of the k lines of data. This was performed by calculated the average of voxel intensity, beginning 5 voxels outside the head extending laterally to include 95 voxels. This was then averaged over the k-lines of data.

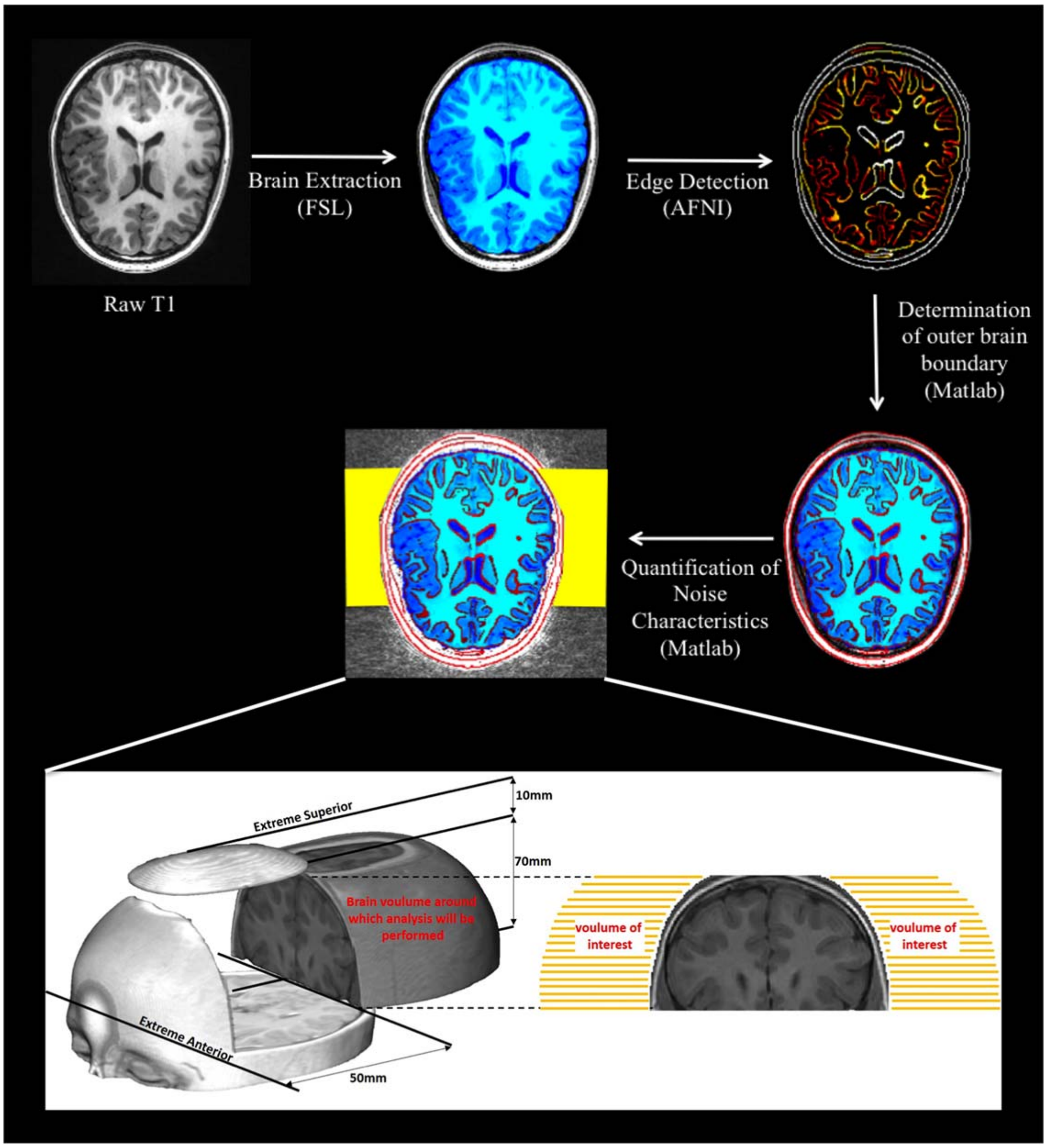
Raw T1
Brain Extraction (FSL)
Edge Detection (AFNI)
Determination of outer brain boundary (Matlab)
Quantification of Noise Characteristics (Matlab)
Extreme Superior
10mm
70mm
Brain voulume around which analysis will be performed
voulume of interest
voulume of interest
Extreme Anterior
50mm

FIGURE 1 Processing steps for the automated quality assessment algorithms. The algorithm begins with the T_1 nii image and uses a combination of FSL, AFNI, and in-house Matlab programs. The region of interest used in the three different quality assessment algorithms is shown in the insert [Color figure can be viewed at wileyonlinelibrary.com]

3.3.3 | Gradient

Head movement during scanning in the frequency domain results in smoothing in the spatial domain, as movement is akin to convolution with the smoothing kernel dependent on the amount of movement. Thus, the effect of movement is especially prominent along regions where there is a sharp change in contrast. Considering that the optimal high-quality image would show a strong edge effect resembling a step function, as the intensity of the image moves from the edge of the head to the air outside the head. No movement could be considered a delta function, such that the convolution of the step function with the delta function would return a step function, or a sharp contrast between the head/air interface. Movement, however, means that the step function at the edge of the head is convolved with a blurred, Gaussian-like waveform that has greater blurring dependent on movement. One way to measure is to deconvolve the step function with the actual image, to determine the waveform associated with smoothing. However, smoothing has a dual effect along the air/head interface. The peak MR signal resulting from the skin surrounding the skull will decrease in intensity and second, there is blurring radiating outward from the air/head border. Thus, the third approach measures the consequence of smoothing by measuring the gradient from the edge of the head to five voxels outside the head. This approach provides an estimate of the point-spread function as a result of blurring from poor image quality.

4 | STATISTICAL ANALYSES

All statistical analyses were performed using the R statistics package (version 3.1.2). To compare the automated Q/A tool with visual inspection, which is considered the "gold-standard," we utilized receiver operating characteristics (ROC) curves using the R package "*pROC*." The ROC curve and the area under the curve (AUC) were calculated within each cohort. We used the R package "*caret*" for calculating the positive and negative predictive value plots. Finally, we utilized Pearson correlation coefficients to compare the automated Q/A metric with FreeSurfer-derived measures of cortical thickness in the frontal, temporal, parietal, and occipital lobes.

5 | RESULTS

5.1 | Demographics

See Table 1 for the demographic information for each cohort. The age range of all three cohorts spanned from 4.4 to 77.6 years of age. Three of the cohorts consisted only of pediatric populations, including the GUSTO cohort that included preschool children (mean age 4.59 years), and the two Generation R waves (mean age of 7.9 and 10.1 years for Waves 1 and 2, respectively). While the NHGRI primarily involved children and adolescents, 32% of the sample included participants older than 25 years of age.

5.2 | Manual quality assessment

To generate ROC curves, a dichotomous measure was created for those that pass or fail the Q/A ratings (QA-pass and QA-fail). Descriptive measures of the usable and unusable scans from the visual inspection are shown in Table 1. For the Generation R Study, the QA-fail scans included those with unusable or poor quality ratings. This resulted in 922 QA-pass and 143 QA-fail scans for wave 1 and 3,559 QA-pass and 381 QA-fail scans for wave 2. The NHGRI cohort were rated by two independent trained researchers with a scale between one (excellent) and 4 (unusable). A mean of the two raters was used and scores greater than 2 were considered as QA-fail scans. Because scanning preschool children is extremely challenging, the GUSTO study performed multiple T_1-weighted images on the children, with a total of 811 scans for 252 individuals. Of these scans, 368 of 443 were considered usable (44.8%).

5.3 | Automated quality assessment performance

The ROC curves for the Fourier, integral, and gradient approaches for both the first and second waves within Generation R Waves 1 and 2 are shown in Figure 2. For Waves 1 and 2, the gradient approach performed the best with both imaging waves having an area under the curve (AUC) of 0.95. The integral approach also performed quite well, having an AUC for the first and second waves of 0.90 and 0.92 respectively. The Fourier approach performed the worst of the three algorithms, with an AUC of 0.77 for Wave 1 and 0.62 for Wave 2. Because the narrow field of view used in images collected from the GUSTO study, neither the integral nor the Fourier approach could be applied. Thus, as the gradient approach provided the best results of the three approaches, we used this algorithm for both the GUSTO and NHGRI cohorts. The AUC for the NHGRI cohort using the gradient approach was 0.88 and 0.82 for the GUSTO cohort (Figure 3). Graphs of the positive- and negative-predictive values for each of the four cohorts are shown in Figure 4.

We also assessed whether the gradient automated Q/A metric could predict visual FreeSurfer quality ratings. Of the 1,070 scans for the first neuroimaging wave of the Generation R Study, five scans were excluded as they failed the FreeSurfer pipeline. Of the 1,065 scans that were constructed, 143 were manual Q/A-fail and 922 were manual Q/A-pass scans. The AUC results for Wave 1 of the Generation R Study were 0.91. The second wave of the Generation R study had a total of 3,973 scans that could be reconstructed, of which 3,234 were rated as useable quality and 689 as unsuable. The AUC results for Wave 2 of the Generation R Study were 0.77. Visual inspection of the NHGRI FreeSurfer constructions rendered two different metrics, the segmentation (internal) and surfaces (external) ratings. The AUC for the NHGRI was 0.73 and 0.66 for the internal and external ratings, respectively. Finally, the AUC for the automated gradient Q/A metric and the GUSTO FreeSurfer generated scans was 0.78.

During the second wave of MRI data collection in the Generation R Study, if the technician found that the scan was of poor quality, then the T_1-weighted image was repeated. Thus, with an initially poor rated scan, we assessed the utility of obtaining a second scan. Figure 5 displays a box plot of the gradient algorithm for the first, second, and best rated scans. A paired *t* test of the automated Q/A value for the first and second scan was highly significant ($t = 19.1$, df = 379, $p < 2.2\text{e-}16$), with the second scan being notably better than the first.

5.4 | Relationship between the automated quality with age and sex

Linear regression showed a significant relationship between the automated Q/A measure and age in both the first ($p = .0001$) and second waves ($p = 5.9 \times 10^{-9}$) of the Generation R Study, and in the NHGRI study ($p = 7.3 \times 10^{-12}$). In all three groups, increased age was associated with better quality data. There was no significant relationship between the automated Q/A metric and age in the GUSTO study. However, both the GUSTO ($p = .02$) and the first wave of the Generation R Study ($p = 6.2 \times 10^{-5}$) showed a significant relationship between sex and the automated Q/A measure, with girls having less movement in the scanner.

5.5 | Automated quality assessment and FreeSurfer derived cortical thickness

For each of the four groups, the relationship between the automated Q/A metric and regional cortical thickness and surface area measures was evaluated. This was performed using MATLAB and by calculating Pearson correlation coefficients between measures of mean cortical thickness and surface area of the frontal, parietal, temporal, and

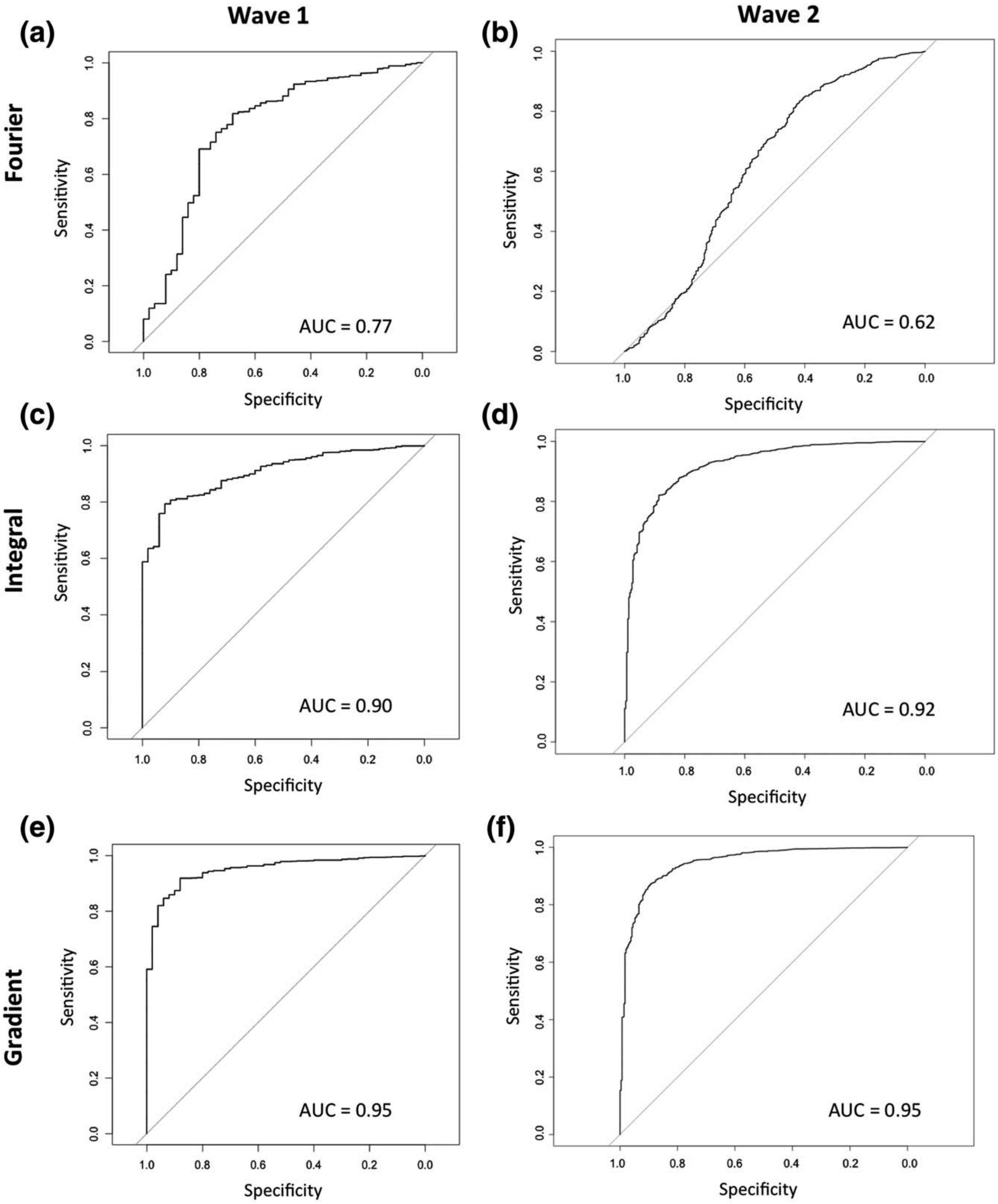

FIGURE 2 Receiver operator characteristics for the three automated quality assessment algorithms: (a) Area under the curve for the Fourier algorithm for Wave 1. (b) Area under the curve for the Fourier algorithm for Wave 2. (c) Area under the curve for the Integral algorithm for Wave 1. (d) Area under the curve for the Integral algorithm for Wave 2. (e) Area under the curve for the Gradient algorithm for Wave 1. (f) Area under the curve for the Gradient algorithm for Wave 2

occipital lobes and the automated Q/A measure. This analysis was performed while iteratively removing a scan with the lowest automated Q/A measure (poorer quality) and then recalculating the correlation coefficient. The results of these analysis are shown in Figure 6 for cortical thickness and Figure 7 for surface area. As cortical thickness and surface area are also associated with age, and for comparability

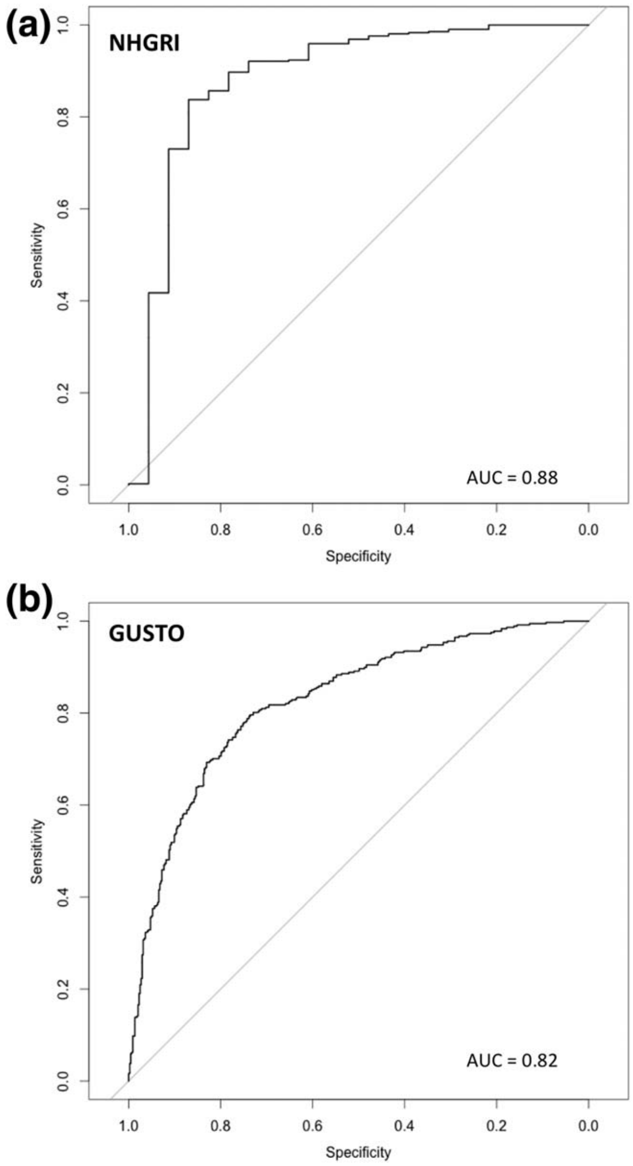

FIGURE 3 Receiver operator characteristics for the NHGRI and GUSTO cohorts

between the studies, data from the NHGRI is split into a group younger than 13 years of age.

Finally, we also assessed whether there was a relationship between the automated Q/A metric and FreeSurfer metrics even with the highest rated quality images. To do this, we used the ordered categorical ratings and selected only those scans that were rated the best quality within each of the four cohorts. This resulted in low to moderate correlations, depending on the cohort, with both cortical thickness and surface area (Table 2). The relationship was highest for cortical thickness measures in wave I of the Generation R Cohort, and surface area in the GUSTO cohort.

6 | DISCUSSION

Using two large neuroimaging waves of a large population-based study of child development, we developed and tested three different algorithms to automatically measure the quality of raw T_1-weighted images. Of these three approaches, we found that the optimal approach was measuring the gradient between the edge of the head and the noise outside the head. However, the algorithm that calculated the integral of the noise outside the head was nearly equivalent. The Fourier algorithm, which evaluated spectral patterns of noise radiating away from the head was the least predictive. We tested the gradient approach using two separate cohorts and found relatively high predictive values with the manual ratings. Furthermore, we found that not only can automated Q/A algorithms provide accurate ratings of raw T_1 images, but these measures can also provide some prediction of the quality of postprocessed images. In addition, we found that when scanning school age children, if the initial scan is of poor quality, repeating the scan is worthwhile as there is a good chance that the second scan will be of better quality than the first. Finally, we found that even after excluding large numbers of children due to movement, and even within the best rated scans, a small to moderate correlation remained between raw image quality and FreeSurfer derived measures of cortical thickness and surface area, although with some mixed results in the four different cohorts.

Noise characteristics that can influence scan quality largely fall into two different categories: machine-related and subject-related noise. The rapid advancement in MR technologies has dramatically reduced machine-related noise, although the regular use of phantoms is important to monitor scanner stability and geometric distortions over time (Bourel, Gibon, Coste, Daanen, & Rousseau, 1999, Friedman and Glover, 2006, Maikusa et al., 2013). In addition, major upgrades to MR hardware or software, while not considered noise, can influence image quality, and is especially important to consider in longitudinal studies. While major sources of subject-related noise include ghosting, aliasing, chemical shifts, and flow artifacts (Hahn et al., 1988, Mirowitz, 1999, Mortamet et al., 2009), the major challenges associated with pediatric neuroimaging involve motion related artifacts (Raschle et al., 2009, Dean et al., 2014).

Motion that occurs with data acquisition in the frequency domain can have two major effects in the spatial domain. First, movement that is periodic in nature (i.e., respiratory or cardiac related) occurs over the entire imaging sequence, and thus is observed as ghosting artifacts present in the spatial domain along the phase encoding direction (Saloner, 1999). Second, aperiodic movement of the participants, such as "wiggly" children, typically occurs between the pulse excitation and the echo, resulting in spin incoherence of the phase at the time of the echo (Barish and Jara, 1999). This incoherence, following a Fourier transform, results in increased noise and blurring in the spatial domain. Given these patterns of subject-related noise, our use of three specific algorithms (Fourier, integral, and edge gradient) were applied and tested so as to capture the primary aspects of each of these specific movement-related artifacts.

While visual inspection of images remains crucial, especially for the identification of incidental findings (Jansen et al., 2017), automated measures can provide important quantitative information. In fact, comparing an automated versus visual ratings with images that were manipulated, Gardner et al. (1995) found that visual raters were unable

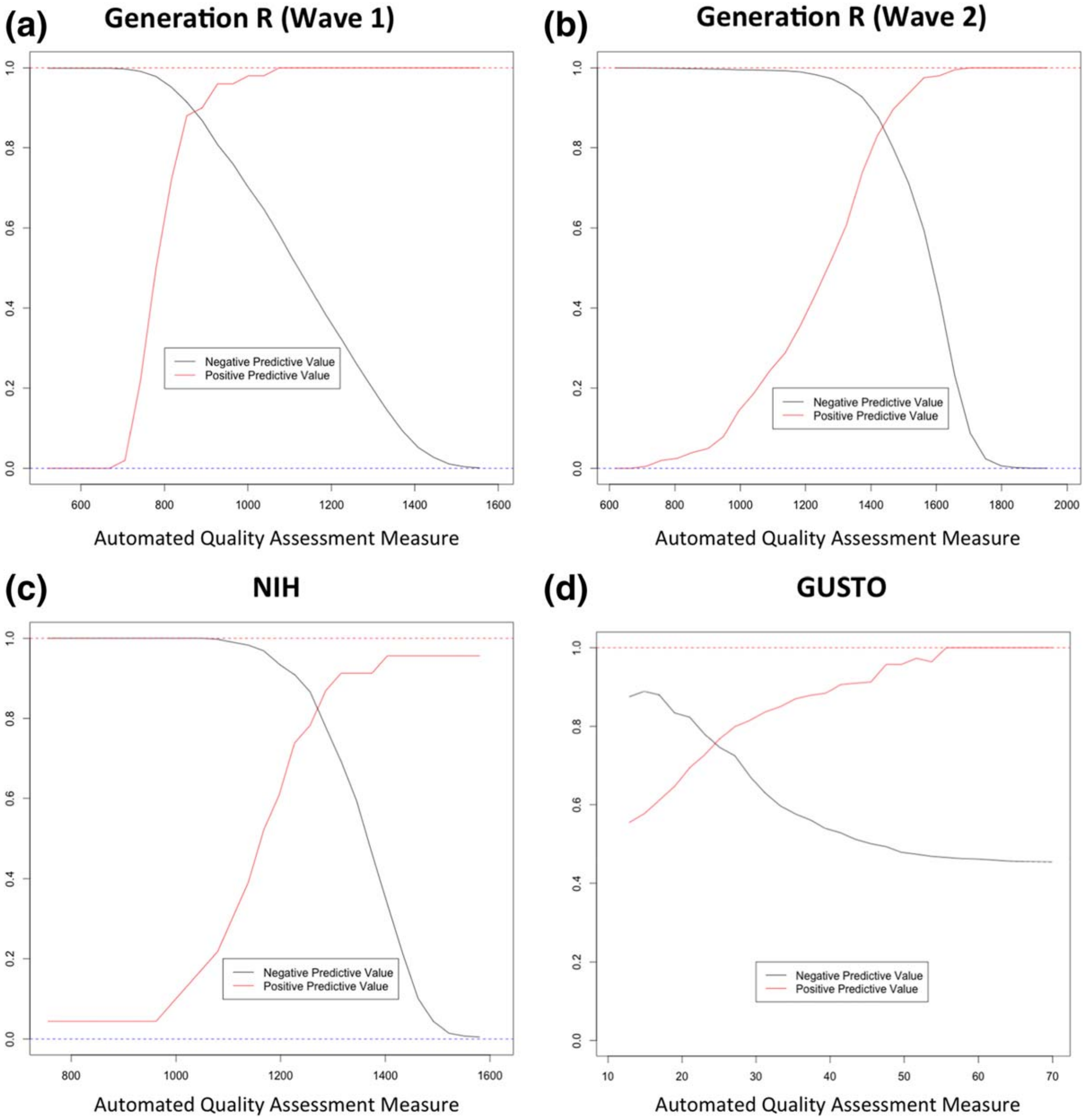

FIGURE 4 Positive predictive value and negative predictive value plots for the two Generation R waves and the NHGRI and GUSTO cohorts [Color figure can be viewed at wileyonlinelibrary.com]

to detect slice thickness increases of 40%, whereas the automated approach was able to quantify even minor changes. Furthermore, for large population-based studies where multiple scans need to be rated, visual inspection of data can be prone to rater-differences (inter-rater reliability) and rater-drift (intra-rater reliability), which is not a problem with automated approaches. There have been several algorithms developed to automatically assess the quality of structural images (Mortamet et al., 2009, Pizarro et al., 2016). Mortamet et al. (2009) measured voxel intensities outside of the head with the hypothesis that artifacts enlarge the noise intensity and causes a right-skew (greater intensity) in the distribution. The authors applied the algorithm to a group of 188 elderly subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008) and found ROC characteristics similar to our gradient and integral approaches (AUC = 0.94). Pizarro et al. (2016) argued that multiple metrics, rather than one global metric, would provide better measures of Q/A and presented findings from an automated structural QA algorithm that extracted multiple features from the brain and surrounding noise and entered these features into a support vector machine (SVM) to classify data quality. They reported an ~80% accuracy with their SVM approach, where the lower accuracy could reflect the multisite nature of their study or alternatively, increased noise due to using multiple brain features. To date, there have been no studies evaluating automated Q/A in children and assessing these metrics directly with outcome measures such as cortical thickness.

Recent functional MRI studies have demonstrated that even small amounts of subject motion in children can affect the quantification of connectivity metrics (Power et al., 2012). To reduce head motion and also to get children acclimatized to the scanner and the environment, many research studies train subjects with a mock scanning session. In

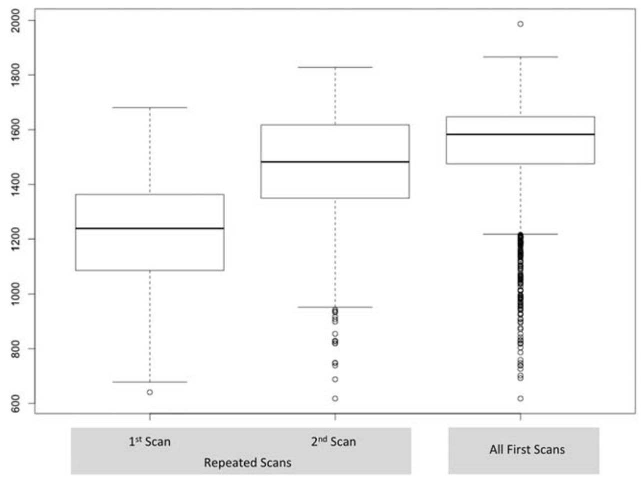

FIGURE 5 Boxplot demonstrating the improvement in obtaining a repeat scan when the fist T_1 image is poor

spite of these efforts, motion in children cannot be completely alleviated and continues to pose a problem in the quality of the scans. Our study provides important information to this ongoing discussion of the role of quality and MR derived metrics by showing that the correlation between structural imaging Q/A and cortical thickness is present in samples not only when the poorest quality images are included, but when many scans that passed Q/A were included. This finding supports using automated algorithms to assess for relationships with reconstruction metrics, and when such relationship exist, metrics from an automated Q/A algorithm should be used as a covariate to adjust for small differences in movement.

The strengths of the study include the large sample in four different groups of children and four different scanners, with the samples drawn from the general population. We tested and compared three different algorithms for quality assessment of structural images within the Generation R Study and tested the best performing metric in two independent samples. Finally, it is a strength that we also compared these findings to postprocessing streams, to assess for downstream effects. There are also several weaknesses of the study. First, we did not measure heart rate, respiratory rate, eye tracking, and external fiduciary markers to more precisely quantify the different forms of artifacts. Such an approach would be beneficial to assess which types of subject-dependent noise have the greatest influence. In addition, the cutoff that we used for useable versus not usable scans was different at each of the three sites. However, these differences provide a greater "real-world" application for our findings. Since the algorithm was first optimized and tested within the Generation R Study, it is possible that it was more "tuned" for the gradient sequences used for the Rotterdam site. Thus, it may be possible to tweak the algorithm to show improvement within each site. Finally, although we performed the automated Q/A algorithm on four different scanners, three of them were GE scanners (one was a wide-bore scanner) and it is possible other different vendors and models may have internal software, such as edge sharpening algorithms, that would make the gradient approach less accurate. Thus, it is important to test the algorithm in a wide variety of MR

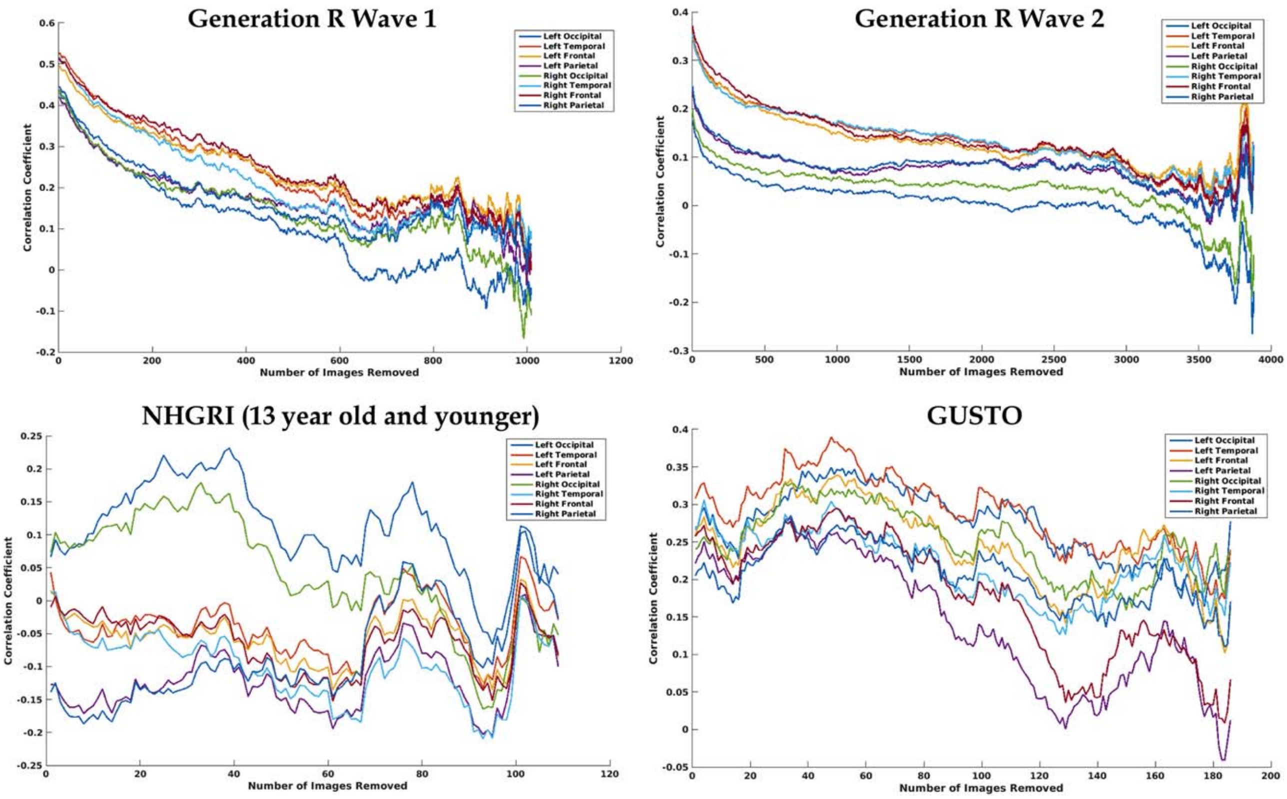

FIGURE 6 The association between the automated quality assessment algorithms and FreeSurfer derived cortical thickness from the four major lobes [Color figure can be viewed at wileyonlinelibrary.com]

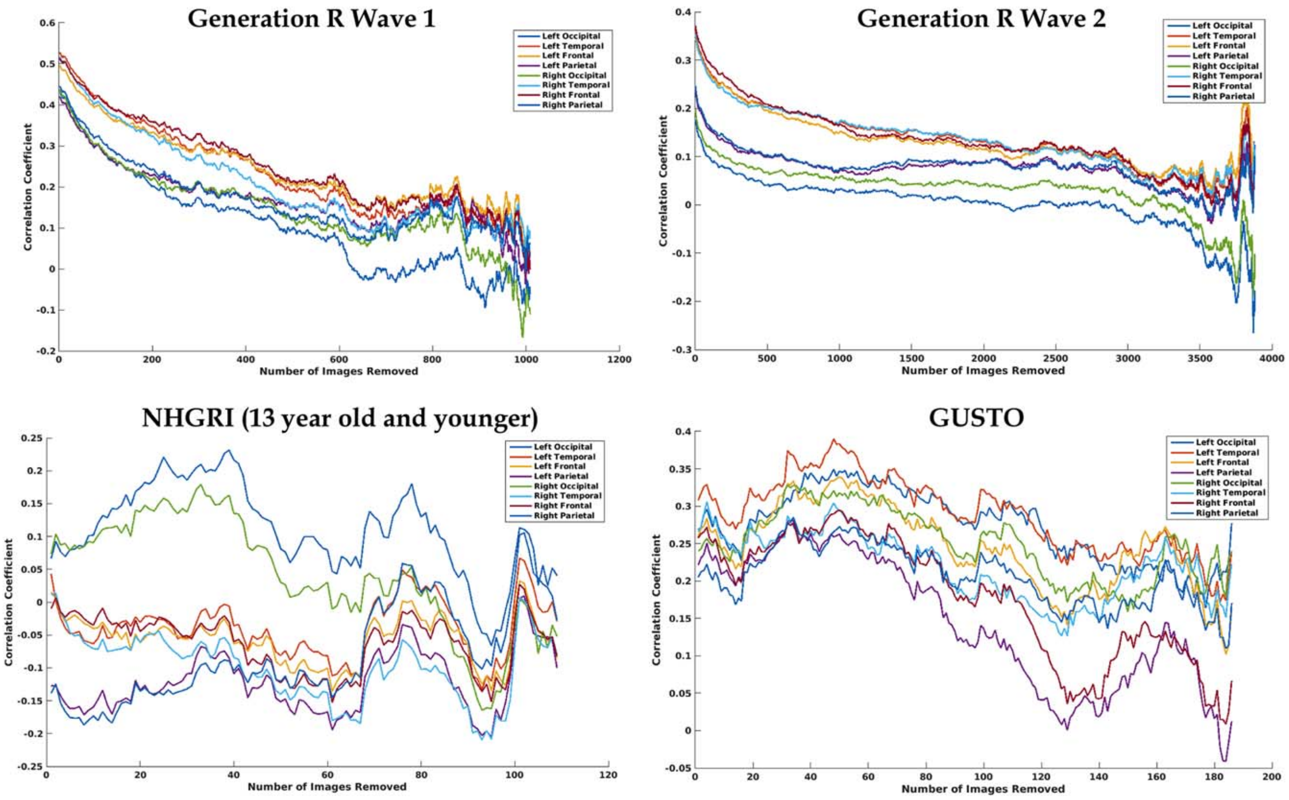

FIGURE 7 The association between the automated quality assessment algorithms and FreeSurfer derived surface area from the four major lobes [Color figure can be viewed at wileyonlinelibrary.com]

systems and determine whether normalization of the different distributions of the automated Q/A metric would allow for harmonization between scanners. Since the distributions of the automated Q/A metrics were normally distributed for each scanner, determining scanner-specific thresholds may be easily accomplished by selecting a cutoff associated with a specific *z*-value for each distribution.

In conclusion, we designed and tested three different automated approaches to measure the quality of structural MR images. We found that a simple gradient approach, which tapped into the principle of the line-spread function and measured the gradient between the edge of the head and noise outside the head performed slightly better in both waves of pediatric neuroimaging data in the Generation R Study. This

TABLE 2 Pearson correlation coefficients between the automated quality assessment metric and FreeSurfer-based cortical thickness and surface area measures using only the best rated quality scans within each cohort

	Generation R Wave I (*n* = 227)	Generation R Wave II (*n* = 365)	NHGRI (*n* = 106)	GUSTO (*n* = 122)
Cortical thickness				
Left frontal	0.30	0.14	0.15	−0.02
Right frontal	0.32	0.27	0.12	−0.04
Left temporal	0.30	0.16	0.19	0.13
Right temporal	0.32	−0.01	0.17	0.08
Left parietal	0.04	0.11	0.09	0.07
Right parietal	0.06	0.08	0.16	0.06
Left occipital	−0.07	0.15	0.05	−0.06
Right occipital	0.03	0.04	0.08	0.03
Surface area				
Left frontal	0.12	0.13	−0.13	0.18
Right frontal	0.12	0.13	−0.14	0.23
Left temporal	0.06	0.13	−0.15	0.27
Right temporal	0.01	0.10	−0.17	0.25
Left parietal	0.01	0.10	−0.25	0.26
Right parietal	−0.02	0.10	−0.20	0.15
Left occipital	0.09	0.01	−0.01	0.21
Right occipital	0.10	0.07	0.01	0.19

algorithm was then tested in two separate cohorts (NHGRI and GUSTO) and demonstrated that the predictive value for automated Q/A rating was, while less than in the Generation R Study, quite good. In addition, we demonstrated that in school age children, there is utility in repeating the structural scan if the first scan has poor quality, as the chances are high that the second scan is better than the first. During the scanning session, if we saw that the scan quality was poor, we explained to the children that the scan was blurry because of movement and would need to be repeated. We then kindly encouraged the children to remain as still as possible. This may have had a positive effect, as the second scan was on average considerably better. We also found that the quality of the raw T1 image has good predictive power for the quality of the FreeSurfer surface reconstruction. Finally, we found results similar to those in fMRI studies, that even small movements can have an influence on FreeSurfer-derived cortical reconstruction measures. Thus, our findings suggest that automated measures of head movement can serve as a helpful adjunct to visual inspection. Further research should be directed to asesss whether such automated Q/A metrics should be used as covariates in structural MRI analyses, similar to regressing motion from fMRI data, or to provide additional information in selecting thresholds to exclude participants based on poor image quality.

ACKNOWLEDGMENTS

Neuroimaging studies within the Generation R are supported through the Netherlands Organization for Health Research and Development (ZonMw TOP 40–00812-98–11021), the European Community's 7th Framework Programme (FP7/2008–2013) under grant agreement 212652 (NUTRIMENTHE), and the Stichting Sophia Kinderziekenhuis Fonds. The Generation R Study is conducted by the Erasmus Medical Center in close collaboration with the Erasmus University Rotterdam, the Municipal Health Service in the Rotterdam area, the Rotterdam Homecare Foundation, Rotterdam and the Stichting Trombosedienst & Artsenlaboratorium Rijnmond (STAR-MDC), Rotterdam. We gratefully acknowledge the contribution of children and parents, general practitioners, hospitals, midwives, and pharmacies in Rotterdam. The Generation R Study is made possible by financial support from the Erasmus Medical Center, Rotterdam, the Erasmus University Rotterdam, the Netherlands Organization for Scientific Research (NWO), the Ministry of Health, Welfare and Sport, and the Ministry of Youth and Families. I would also like to thank Pierre Olivier Quirion for his assistance in assisting with the conversion of my Matlab code to Python during the 2016 OHBM Hackathon in Vancouver, BC.

CONFLICTS OF INTEREST

None of the authors have any conflicts of interest associated with this study.

ORCID

Tonya White http://orcid.org/0000-0003-0271-1896
Anqi Qiu http://orcid.org/0000-0002-0215-6338

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. *Human Brain Mapping*, *37*, 2385–2397.

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 Transactions on Medical Imaging*, *16*, 903–910.

Backhausen, L. L., Herting, M. M., Buse, J., Roessner, V., Smolka, M. N., & Vetter, N. C. (2016). Quality control of structural MRI images applied using FreeSurfer-A hands-on workflow to rate motion artifacts. *Frontiers in Neuroscience*, *10*, 558.

Barish, M. A., & Jara, H. (1999). Motion artifact control in body MR imaging. *Magnetic Resonance Imaging Clinics of North America*, *7*, 289–301.

Blumenthal, J. D., Zijdenbos, A., Molloy, E., & Giedd, J. N. (2002). Motion artifact in magnetic resonance imaging: Implications for automated analysis. *NeuroImage*, *16*, 89–92.

Bourel, P., Gibon, D., Coste, E., Daanen, V., & Rousseau, J. (1999). Automatic quality assessment protocol for MRI equipment. *Medical Physics*, *26*, 2693–2700.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*, 179–194.

Dean, D. C., 3rd, Dirks, H., O'muircheartaigh, J., Walker, L., Jerskey, B. A., Lehman, K., … Deoni, S. C. (2014). Pediatric neuroimaging using magnetic resonance imaging during non-sedated sleep. *Pediatric Radiology*, *44*, 64–72.

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*, 968–980.

Ducharme, S., Albaugh, M. D., Nguyen, T. V., Hudziak, J. J., Mateos-Perez, J. M., Labbe, A., … Karama, S. Brain Development Cooperative, G. (2016). Trajectories of cortical thickness maturation in normal brain development–The importance of quality control procedures. *NeuroImage*, *125*, 267–279.

El Marroun, H., Schmidt, M. N., Franken, I. H., Jaddoe, V. W., Hofman, A., van der Lugt, A., … White, T. (2014). Prenatal tobacco exposure and brain morphology: A prospective study in young children. *Neuropsychopharmacology*, *39*, 792–800.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*, 774–781.

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11050–11055.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999a). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*, 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*, 272–284.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D. H., … Dale, A. M. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, *14*, 11–22.

Friedman, L., & Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *Journal of Magnetic Resonance Imaging*, *23*, 827–839.

Gardner, E. A., Ellis, J. H., Hyde, R. J., Aisen, A. M., Quint, D. J., & Carson, P. L. (1995). Detection of degradation of magnetic resonance (MR) images: Comparison of an automated MR image-quality analysis system with trained human observers. *Academic Radiology*, *2*, 277–281.

Hahn, F. J., Chu, W. K., Coleman, P. E., Anderson, J. C., Dobry, C. A., Imray, T. J., . . . Lee, S. H. (1988). Artifacts and diagnostic pitfalls on magnetic resonance imaging: A clinical review. *Radiologic Clinics of North America*, *26*, 717–735.

Jack, C. R., Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., . . . Weiner, M. W. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*, 685–691.

Jaddoe, V. W., Mackenbach, J. P., Moll, H. A., Steegers, E. A., Tiemeier, H., Verhulst, F. C., . . . Hofman, A. (2006). The Generation R Study: Design and cohort profile. *European Journal of Epidemiology*, *21*, 475–484.

Jansen, P. R., van der Lugt, A., & White, T. J. H. (2017). Incidental findings on brain imaging in the general pediatric population. *New England Journal of Medicine*, *377*, 1593–1595.

Maikusa, N., Yamashita, F., Tanaka, K., Abe, O., Kawaguchi, A., Kabasawa, H., . . . Iwatsubo, T. Japanese Alzheimer's Disease Neuroimaging, I. (2013). Improved volumetric measurement of brain structure with a distortion correction procedure using an ADNI phantom. *Medical Physics*, *40*, 062303.

Mirowitz, S. A. (1999). MR imaging artifacts. Challenges and solutions. *Magnetic Resonance Imaging Clinics of North America*, *7*, 717–732.

Mortamet, B., Bernstein, M. A., Jack, C. R., Jr., Gunter, J. L., Ward, C., Britson, P. J., . . . Krueger, G. Alzheimer's Disease Neuroimaging, I. (2009). Automatic quality assessment in structural brain magnetic resonance imaging. *Magnetic Resonance in Medicine*, *62*, 365–372.

Pizarro, R. A., Cheng, X., Barnett, A., Lemaitre, H., Verchinski, B. A., Goldman, A. L., . . . Mattay, V. S. (2016). Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm. *Frontiers in Neuroinformatics*, *10*,

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.

Raschle, N. M., Lee, M., Buechler, R., Christodoulou, J. A., Chang, M., Vakil, M., . . . Gaab, N. (2009). Making MR imaging child's play - pediatric neuroimaging protocol, guidelines and procedure. *Journal of Visualized Experiments.*

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, *107*, 107–115.

Saloner, D. (1999). Flow and motion. *Magnetic Resonance Imaging Clinics of North America*, *7*, 699–715.

Satterthwaite, T. D., Wolf, D. H., Loughead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., . . . 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.

Segonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, *22*, 1060–1075.

Segonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Transactions on Medical Imaging*, *26*, 518–529.

Soh, S. E., Lee, S. S. M., Hoon, S. W., Tan, M. Y., Goh, A., Lee, B. W., . . . van Bever, H. P. S. (2012). The methodology of the GUSTO cohort study: A novel approach in studying pediatric allergy. *Asia Pacific Allergy*, *2*, 144–148.

Tiemeier, H., Velders, F. P., Szekely, E., Roza, S. J., Dieleman, G., Jaddoe, V. W., . . . Verhulst, F. C. (2012). The Generation R Study: A review of design, findings to date, and a study of the 5-HTTLPR by environmental interaction from fetal life onward. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*, 1119–1135. e1117.

Van Dijk, K. R., Sabuncu, M. R., & Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage*, *59*, 431–438.

White, T., El Marroun, H., Nijs, I., Schmidt, M., van der Lugt, A., Wielopolki, P. A., . . . Verhulst, F. C. (2013). Pediatric population-based neuroimaging and the Generation R Study: The intersection of developmental neuroscience and epidemiology. *European Journal of Epidemiology*, *28*, 99–111.

White, T., Muetzel, R. L., El Marroun, H., Blanken, L. M. E., Jansen, P., Bolhuis, K., . . . Tiemeier, H. (2017). Paediatric population neuroimaging and the Generation R Study: The second wave. *European Journal of Epidemiology*.

How to cite this article: White T, Jansen PR, Muetzel RL, et al. Automated quality assessment of structural magnetic resonance images in children: Comparison with visual inspection and surface-based reconstruction. *Hum Brain Mapp*. 2017;00:1–14. https://doi.org/10.1002/hbm.23911