Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines

Oscar Esteban[1], Ross W Blair[1], Dylan M Nielson[2], Jan C Varada[3], Sean Marrett[3], Adam G Thomas[2], Russell A Poldrack[1], and Krzysztof J Gorgolewski[1]

1. Dept. of Psychology, Stanford University, CA, USA.
2. Data Science and Sharing Team, National Institute of Mental Health, Bethesda, MD, USA.
3. Functional MRI Facility, National Institute of Mental Health, Bethesda, MD, USA .
corresponding author(s): Oscar Esteban (phd@oscaresteban.es)

Abstract

The neuroimaging community is steering towards increasingly large sample sizes, which are highly heterogeneous because they can only be acquired by multi-site consortia. The visual assessment of every imaging scan is a necessary quality control step, yet arduous and time-consuming. A sizeable body of evidence shows that images of low quality are a source of variability that may be comparable to the effect size under study. We present the MRIQC WebAPI, an open crowdsourced database that collects image quality metrics extracted from MR images and corresponding manual assessments by experts. The database is rapidly growing, and currently contains over 100,000 records of image quality metrics of functional and anatomical MRIs of the human brain, and over 200 expert ratings. The resource is particularly designed for researchers to share image quality metrics and annotations that can readily be reused in training human experts and machine learning algorithms. The ultimate goal of the MRIQC WebAPI is to allow the development of fully automated quality control tools that outperform expert ratings in identifying subpar images.

Background & Summary

Ensuring the quality of neuroimaging data is a crucial initial step for any image analysis workflow because low quality images may obscure the effects of scientific interest[1-4]. Most approaches use manual quality control (QC), which entails screening every single image of a dataset individually. However, manual QC suffers from at least two problems: unreliability and time-consuming nature for large datasets. Unreliability creates great difficulty in defining objective exclusion criteria in studies, and stems from intrinsically large intra-rater and inter-rater variabilities[5]. Intra-rater variability derives from aspects such as training, subjectivity, varying annotation settings and protocols, fatigue or bookkeeping errors. The difficulty in calibrating between experts lies at the heart of inter-rater variability. In addition to the need for objective exclusion criteria, the current neuroimaging data deluge makes the manual QC of every MRI scan impractical. For these reasons, there has been great interest in automated QC[5-8], which is progressively gaining attention[9-16] with the convergence of machine learning solutions. Early approaches[5-8] to objectively estimate image quality have employed metrics ("image quality metrics", IQMs) that quantify variably interpretable aspects of image quality[9-15] (e.g. summary statistics of image intensities, signal-to-noise ratio, coefficient of joint variation, Euler angle, etc.). The approach has been shown sufficiently reliable in single-site samples[8,11-15], but it does not generalize well to new images acquired at sites unseen by the decision algorithm[9]. Decision algorithms do not generalize to new datasets because the large between-site variability as compared to the within-site variability of features poses a challenging normalization problem, similar to "batch-effects" in genomic analyses[19]. Additional pitfalls limiting fully automated QC of MRI relate to the small size of databases that include quality annotations, and the unreliability of such annotations (or

"*labels noise*"). As described previously, rating the quality of every image in large databases is an arduous, unreliable, and costly task. The convergence of limited size of samples annotated for quality, and the labels noise preclude the definition of normative, standard values for the IQMs that work well for any dataset, and also, the generalization of machine learning solutions. Keshavan et al.[16] have recently proposed a creative solution to the problem of visually assessing large datasets. They were able to annotate over 80,000 bidimensional slices extracted from 722 brain 3D images using BraindR, a smart phone application for crowdsourcing. They also proposed a novel approach to the QC problem by training a convolutional neural network using BraindR ratings, with excellent results (AUC 0.99). Their QC tool performed as well as MRIQC[9] (which uses IQMs and a random forests classifier to decide which images should be excluded) on their single-site dataset. By collecting several ratings per screened entity, they were able to effectively minimize the labels noise problem by averaging expert ratings. As limitations to their work we would count the use of 2D images for annotation and the lack of testing their trained network in out-of-sample examples. In sum, automating QC requires large datasets collected across sites, and rated by many individuals in order to ensure generalizability.

The MRIQC Web-API automatically collects manual quality ratings from users of the MRIQC software. As such, it provides a unique platform to address the issues raised above. First, within fourteen months we have collected over 50,000 and 60,000 records of anatomical and functional IQMs, respectively (Figure 1). These IQMs are extracted and automatically submitted (unless the user opts out) with MRIQC (Figure 2), including both publicly available images and private datasets. Second, we leverage the maximized efficiency of MRIQC's reports in assessing individual 3D images with a simplified interface that allows experts to submit their ratings with few clicks (Figure 3). This assessment protocol avoids clerical errors from the operator, as ratings as automatically handled and registered. Since the assessments can be posted to the public database, they can be readily reused. In other words, MRIQC users are building a very large database with minimal effort every day. As only the IQMs and manual ratings are crowdsourced (i.e. no MRI data is shared), data collection is not limited to public datasets only; however,image hashes are stored in order to allow identification of matching images. The presented resource is envisioned to train automatic QC tools and to develop human expert training programs.

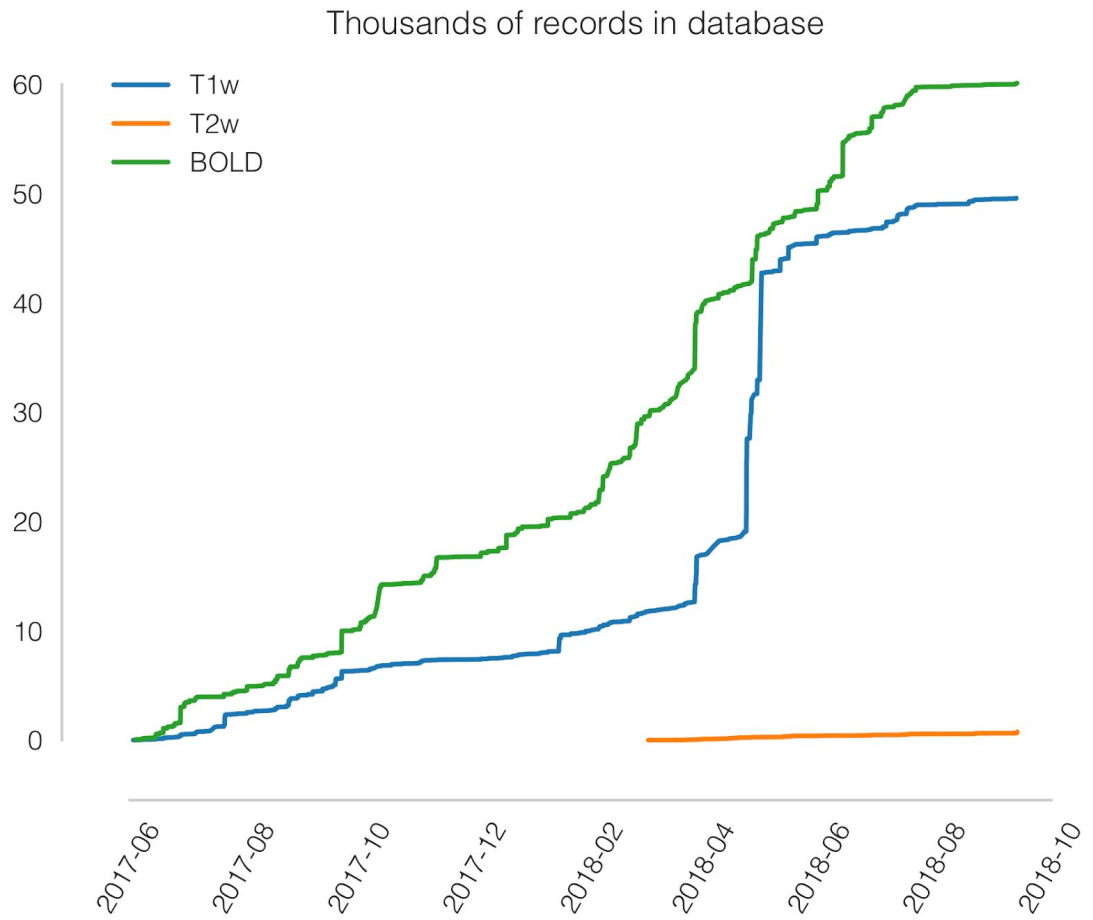

Figure 1. A rapidly growing MRI quality control knowledge base. Evolution of the cumulative number of unique records, per modality.

Methods

Here we describe an open database which collects both IQM vectors extracted from functional and anatomical MRI scans, along with quality assessments done by experts based on visual inspection of images. Although it was envisioned as a lightweight web-service tailored to MRIQC, the database is able to receive new records from any other software, provided they are able to correctly query the API (application programming interface).

Data generation and collection workflow

The overall framework involves the following workflow (summarized in Figure 2):

1. Execution of MRIQC and submission of IQMs: T1w, T2w and BOLD images are processed with MRIQC, which computes a number of IQMs (see section Technical Validation). The IQMs are collated in a JSON record, which MRIQC automatically submits to a Representational state transfer (REST) or RESTful endpoint of the Web-API. Users can opt-out if they do not wish to share their IQMs.
2. JSON records are received by the endpoint, validated and stored in the database. Each record includes the vector of IQMs, a unique identifier generated from the actual data array of the original image, and additional anonymized metadata and provenance.
3. Visualization of the individual reports: MRIQC generates dynamic HTML (hypertext markup language) reports that speed up the visual assessment of each image of the dataset by the expert. MRIQC version 0.12.2 includes a widget (see Figure 2) that

allows the researcher to assign a quality rating to the image being screened (see Table 3).

4. Crowdsourcing expert quality ratings: the RESTful endpoint receives the quality ratings, which are linked to the original image via their unique identifier.
5. Retrieving records: the database can be queried for records with any HTTP client or via web using our interface: https://mriqc.nimh.nih.gov/. Additionally, a snapshot of the database at the time of writing has been deposited to FigShare[29].

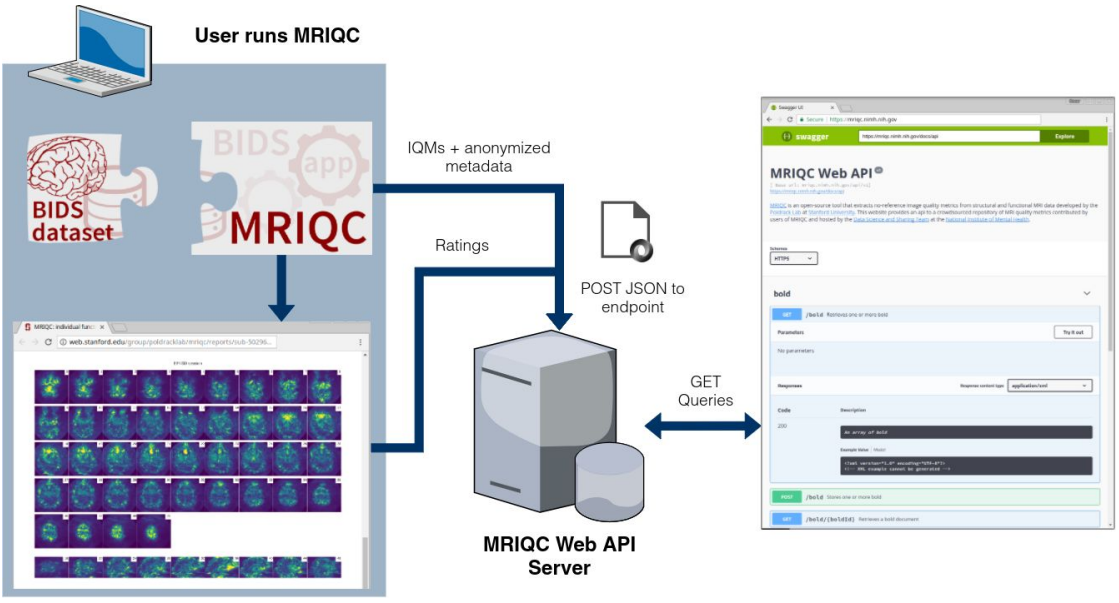

Figure 2. Experimental workflow to generate the database. A dataset is processed with MRIQC. Processing finishes with a POST request to the MRIQC Web API endpoint with a payload containing the image quality metrics (IQMs) and some anonymized metadata (e.g. imaging parameters, the unique identifier for the image data, etc.) in JSON format. Once stored, the endpoint can be queried to fetch the crowdsourced IQMs. Finally, a widget (Figure 3) allows the user to annotate existing records in the MRIQC Web API.

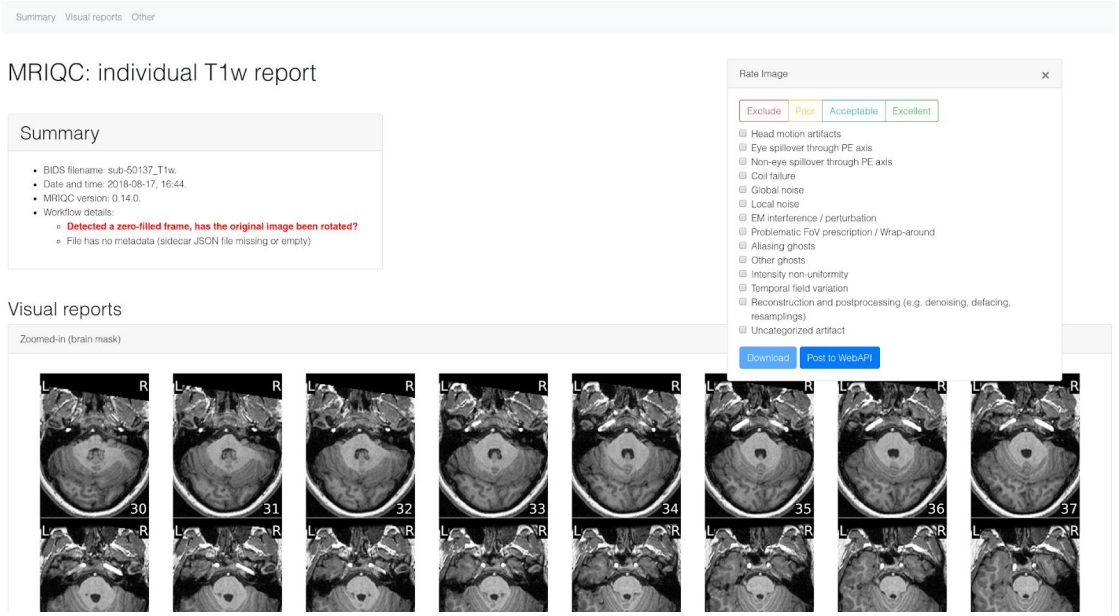

Figure 3. MRIQC visual reports and feedback tool. The visual reports generated with MRIQC include the "Rate Image" widget. After screening of the particular dataset, the expert can assign one quality level (among "exclude", "poor", "acceptable", and "excellent") and also select from a list of MR artifacts typically found in MRI datasets. When the annotation is finished, the user can download the ratings to their local hard disk and submit them to the Web API.

A database to train machines and humans

Primarily, the database was envisioned to address three use-cases:

1. Sampling the distribution of IQMs across all (public and private) datasets, across all scanning sites. Based on this information, researchers can explore questions such as the relationship of particular imaging parameters (e.g. MR scan vendor, or more interestingly, the multi-band acceleration factor or newest fMRI sequences) with respect to the signal-to-noise ratio or the power of N/2 aliasing ghosts.
2. To our knowledge, the community lacks a large database of multi-site MRI annotated for quality that permits the application of machine learning techniques to automate QC. As Keshavan et al. have demonstrated, minimizing the time cost and fatigue load along with the elimination of bookkeeping tasks in the quality assessment of individual MR scans enables collection and annotation of massive datasets.
3. With the new rating feedback feature of MRIQC, institutions can train their experts and compare their assessments against the existing quality annotations. Programs for training experts on quality assessment can be designed to leverage the shared knowledge on the database.

Code availability

The MRIQC Web API is available under the Apache-2.0 license. Source code is accessible through GitHub (https://github.com/poldracklab/mriqcwebapi).

MRIQC is one possible client to generate IQMs and submit ratings feedback. It is available under the BSD 3-clause license. Source code is publicly accessible through GitHub (https://github.com/poldracklab/mriqc).

Data Records

A full data dump of the database at the time of submission is available at FigShare[29]. The database can be programmatically queried to get all the currently stored records through its RESTful API.

MRIQC reports, generated for all T1w images found in OpenfMRI are available for expert training at https://mriqc.s3.amazonaws.com/index.html#openfmri/.

Technical Validation

MRIQC extends the list of IQMs from the QAP, which was constructed from a careful review of the MRI and medical imaging literature[10]. The technical validity of measurements stored to the database is demonstrated by our previous work[8] on the MRIQC client tool and its documentation website: https://mriqc.readthedocs.io/en/latest/measures.html. Definitions for the anatomical IQMs are given in Table 1, and for functional IQMs in Table 2. Finally, the structure of data records containing the manual QC feedback is summarized in Table 3.

Table 1. Summary table of image quality metrics for anatomical (T1w, T2w) MRI. MRIQC produces a vector of 64 image quality metrics (IQMs) per input T1w or T2w scan. (Reproduced from https://doi.org/10.1371/journal.pone.0184661.t002)

IQMs based on noise measurements	
CJV	The coefficient of joint variation of GM and WM was proposed as objective function by Ganzetti et al.[20] for the optimization of INU correction algorithms. Higher values are related to the presence of heavy head motion and large INU artifacts.
CNR	The contrast-to-noise ratio[21] is an extension of the SNR calculation to evaluate how separated the tissue distributions of GM and WM are. Higher values indicate better quality.
SNR	MRIQC includes the the signal-to-nose ratio calculation proposed by Dietrich et al.[22], using the air background as noise reference. Additionally, for images that have undergone some noise reduction processing, or the more complex noise realizations of current parallel acquisitions, a simplified calculation using the within tissue variance is also provided.
QI_2	The second quality index of Mortamet et al.[8] is a calculation of the goodness-of-fit of a χ^2 distribution on the air mask, once the artifactual intensities detected for computing the QI_1 index have been removed. The description of the QI_1 is found below.
IQMs based on information theory	
EFC	The entropy-focus criterion[23] uses the Shannon entropy of voxel intensities as an indication of ghosting and blurring induced by head motion. Lower values are better.
FBER	The foreground-background energy ratio[10] is calculated as the mean energy of image values within the head relative the mean energy of image values in the air mask. Consequently, higher values are better.
IQMs targeting specific artifacts	
INU	MRIQC measures the location and spread of the bias field extracted estimated by the intensity non-uniformity (INU) correction. The smaller spreads located around 1.0 are better.
QI_1	Mortamet's first quality index[8] measures the amount of artifactual intensities in the air surrounding the head above the nasio-cerebellar axis. The smaller QI_1, the better.
WM2MAX	The white-matter to maximum intensity ratio is the median intensity within the WM mask over the 95% percentile of the full intensity distribution, that captures the existence of long tails due to hyper-intensity of the carotid vessels and fat. Values should be around the interval [0.6, 0.8]
Other IQMs	
FWHM	The full-width half-maximum[24] is an estimation of the blurriness of the image using AFNI's 3dFWHMx. Smaller is better.
ICVs	Estimation of the icv of each tissue calculated on the FSL fast's segmentation. Normative values fall around 20%, 45% and 35% for cerebrospinal fluid (CSF), WM and GM, respectively.
rPVE	The residual partial volume effect feature is a tissue-wise sum of partial volumes that fall in the range [5%-95%] of the total volume of a pixel, computed on the partial volume maps generated by FSL fast. Smaller residual partial volume effects (rPVEs) are better.
SSTATs	Several summary statistics (mean, standard deviation, percentiles 5% and 95%, and kurtosis) are computed within the following regions of interest: background, CSF, WM, and GM.
TPMs	Overlap of tissue probability maps estimated from the image and the corresponding maps from the ICBM nonlinear-asymmetric 2009c template[25].

Table 2. Summary table of image quality metrics for functional (BOLD) MRI. MRIQC produces a vector of 64 image quality metrics (IQMs) per input BOLD scan.

Spatial IQMs	
EFC, FBER, FWHM, SNR, SSTATs (see Table 1)	
IQMs measuring temporal variations	
tSNR	A simplified interpretation of the original temporal SNR definition by Kruger et al.[26]. We report the median value of the tSNR map calculated like
GCOR	Summary of time-series correlation as in[27] using AFNI's @compute_gcor
DVARS	Spatial standard deviation of the data after temporal differencing. Indexes the rate of change of BOLD signal across the entire brain at each frame of data. DVARS is calculated using *Nipype*, after head-motion correction
IQMs targeting specific artifacts	
FD	Framewise Displacement - Proposed by Power et al.[1] to regress out instantaneous head-motion in fMRI studies. MRIQC reports the average FD.
GSR	The Ghost to Signal Ratio[28] estimates the mean signal in the areas of the image that are prone to N/2 ghosts on the phase encoding direction with respect to the mean signal within the brain mask[10]. Lower values are better.
DUMMY	Number of dummy scans - A number of volumes in the beginning of the fMRI time-series identified as nonsteady state.
IQMs from AFNI	
AOR	AFNI's outlier ratio - Mean fraction of outliers per fMRI volume as given by AFNI's 3dToutcount
AQI	AFNI's quality index - Mean quality index as computed by AFNI's 3dTqual

Table 3. Summary table of quality assessment values. Annotations received through the feedback widget are stored in a separate database collecting one rating value and an array of artifacts present in the image.

Expert rating	
Exclude	Assigned to images that show quality defects that preclude any type of processing
Poor	Assigned to images that, although presenting some quality problem, may tolerate some types of processing. For instance, a T1-weighted image that may be used as co-registration reference, but will probably generate biased cortical thickness measurements.
Acceptable	Assigned to images that do not show any substantial issue that may preclude processing
Excellent	Assigned to images without quality issues
Artifacts	
Vector of boolean values corresponding to the following list of possible artifacts found in the image: ● Head motion artifacts ● Eye spillover through PE axis ● Non-eye spillover through PE axis	

- Coil failure
- Global noise
- Local noise
- EM interference / perturbation
- Problematic FoV prescription / Wrap-around
- Aliasing ghosts
- Other ghosts
- Intensity non-uniformity
- Temporal field variation
- Reconstruction and postprocessing (e.g. denoising, defacing, resamplings)
- Uncategorized artifact

Usage Notes

Jupyter notebooks demonstrating exemplary analyses of these data are available in the MRIQC repository (https://github.com/poldracklab/mriqc/tree/master/notebooks).

Acknowledgements

This work was supported by the Laura and John Arnold Foundation, NIH R01 EB020740, NIH 1R24MH114705-01, and NINDS grant 1U01NS103780-01.

Author contributions

OE - Data curation, investigation, software, validation, visualization, writing & editing.
RWB - Conceptualization, investigation, software, validation, visualization, writing & editing.
DMN - Data curation, investigation, infrastructure, software, validation, writing & editing.
JCV - Investigation, infrastructure, software, validation, writing & editing.
SM - Funding acquisition, infrastructure, resources, supervision, writing & editing.
AGT - Funding acquisition, infrastructure, resources, supervision, writing & editing.
RAP - Funding acquisition, resources, supervision, writing & editing.
KJG - Conceptualization, investigation, software, validation, resources, supervision, writing & editing.

Competing interests

The authors declare no competing interests.

References

1. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage. 2012; 59(3):2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018
2. Yendiki A, Koldewyn K, Kakunoori S, Kanwisher N, Fischl B. Spurious group differences due to head motion in a diffusion MRI study. NeuroImage. 2014; 88:79–90. https://doi.org/10.1016/j.neuroimage.2013.11.027
3. Reuter M, Tisdall MD, Qureshi A, Buckner RL, van der Kouwe AJW, Fischl B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. NeuroImage. 2015; 107:107–115. https://doi.org/10.1016/j.neuroimage.2014.12.006
4. Alexander-Bloch A, Clasen L, Stockman M, Ronan L, Lalonde F, Giedd J, et al. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. Human Brain Mapping. 2016; 37(7):2385–2397. https://doi.org/10.1002/hbm.23180
5. Gardner EA, Ellis JH, Hyde RJ, Aisen AM, Quint DJ, Carson PL. Detection of degradation of magnetic resonance (MR) images: Comparison of an automated MR image-quality analysis

system with trained human observers. Academic Radiology. 1995; 2(4):277–281. https://doi.org/10.1016/S1076-6332(05)80184-9

6. Woodard JP, Carley-Spencer MP. No-Reference image quality metrics for structural MRI. Neuroinformatics. 2006; 4(3):243–262. https://doi.org/10.1385/NI:4:3:243

7. Gedamu EL, Collins Dl, Arnold DL. Automated quality control of brain MR images. Journal of Magnetic Resonance Imaging. 2008; 28(2):308–319. https://doi.org/10.1002/jmri.21434

8. Mortamet B, Bernstein MA, Jack CR, Gunter JL, Ward C, Britson PJ, et al. Automatic quality assessment in structural brain magnetic resonance imaging. Magnetic Resonance in Medicine. 2009; 62(2):365–372. https://doi.org/10.1002/mrm.21992

9. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ (2017) MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PLoS ONE 12(9): e0184661. https://doi.org/10.1371/journal.pone.0184661

10. Shehzad Z, Giavasis S, Li Q, Benhajali Y, Yan C, Yang Z, et al. The Preprocessed Connectomes Project Quality Assessment Protocol—a resource for measuring the quality of MRI data. In: Front. Neurosci. Conf. Neuroinformatics. Cairns, Australia; 2015.

11. Pizarro RA, Cheng X, Barnett A, Lemaitre H, Verchinski BA, Goldman AL, et al. Automated Quality Assessment of Structural Magnetic Resonance Brain Images Based on a Supervised Machine Learning Algorithm. Frontiers in Neuroinformatics. 2016; 10. https://doi.org/10.3389/fninf.2016.00052

12. Backhausen LL, Herting MM, Buse J, Roessner V, Smolka MN, Vetter NC. Quality Control of Structural MRI Images Applied Using FreeSurfer-A Hands-On Workflow to Rate Motion Artifacts. Frontiers in Neuroscience. 2016; 10. https://doi.org/10.3389/fnins.2016.00558

13. Fidel Alfaro-Almagro, Mark Jenkinson, Neal K. Bangerter, Jesper L.R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N. Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, Diego Vidaurre, Matthew Webster, Paul McCarthy, Christopher Rorden, Alessandro Daducci, Daniel C. Alexander, Hui Zhang, Iulius Dragonu, Paul M. Matthews, Karla L. Miller, Stephen M. Smith, Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank, NeuroImage. 2018. 166:400-424. https://doi.org/10.1016/j.neuroimage.2017.10.034.

14. Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data Eduard T. Klapwijk, Ferdi van de Kamp, Mara van der Meulen, Sabine Peters, Lara M. Wierenga. bioRxiv 278358; doi: https://doi.org/10.1101/278358

15. 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. 2018;39:1218–1231.https://doi.org/10.1002/hbm.23911

16. Anisha Keshavan, Jason Yeatman, Ariel Rokem. Combining citizen science and deep learning to amplify expertise in neuroimaging. bioRxiv 363382; doi: https://doi.org/10.1101/363382

17. Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Adams P, et al. Harmonization of cortical thickness measurements across scanners and sites. bioRxiv. 2017; p. 148502.

18. Dylan M Nielson, Francisco Pereira, Charles Y Zheng, Nino Migineishvili, John A Lee, Adam G Thomas, Peter A Bandettini. Detecting and harmonizing scanner differences in the ABCD study - annual release 1.0. bioRxiv 309260; doi: https://doi.org/10.1101/309260

19. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010; 11(10):733–739. https://doi.org/10.1038/nrg2825

20. Ganzetti M, Wenderoth N, Mantini D. Intensity Inhomogeneity Correction of Structural MR Images: A Data-Driven Approach to Define Input Algorithm Parameters. Frontiers in Neuroinformatics. 2016; p. 10. doi: 10.3389/fninf.2016.00010.

21. Magnotta VA, Friedman L, Birn F. Measurement of Signal-to-Noise and Contrast-to-Noise in the fBIRN Multicenter Imaging Study. Journal of Digital Imaging. 2006;19(2):140–147. doi: 10.1007/s10278-006-0264-x.

22. Dietrich O, Raya JG, Reeder SB, Reiser MF, Schoenberg SO. Measurement of signal-to-noise ratios in MR images: Influence of multichannel coils, parallel imaging, and reconstruction filters. Journal of Magnetic Resonance Imaging. 2007;26(2):375–385. doi: 10.1002/jmri.20969.
23. Atkinson D, Hill DLG, Stoyle PNR, Summers PE, Keevil SF. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. IEEE Transactions on Medical Imaging. 1997;16(6):903–910. doi: 10.1109/42.650886.
24. Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, et al. Test-retest and between-site reliability in a multicenter fMRI study. Human Brain Mapping. 2008;29(8):958–972. doi: 10.1002/hbm.20440.
25. Fonov V, Evans A, McKinstry R, Almli C, Collins D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage. 2009;47, Supplement 1:S102. doi: 10.1016/S1053-8119(09)70884-5.
26. Krüger et al., Physiological noise in oxygenation-sensitive magnetic resonance imaging, Magn. Reson. Med. 46(4):631-637, 2001. doi:10.1002/mrm.1240.
27. Saad et al. Correcting Brain-Wide Correlation Differences in Resting-State FMRI, Brain Conn 3(4):339-352, 2013, doi:10.1089/brain.2013.0156.
28. Giannelli, M., Diciotti, S., Tessa, C., & Mascalchi, M. (2010). Characterization of Nyquist ghost in EPI-fMRI acquisition sequences implemented on two clinical 1.5 T MR scanner systems: effect of readout bandwidth and echo spacing. Journal of Applied Clinical Medical Physics, 11(4).
29. Esteban, Oscar; Blair, Ross; Nielson, Dylan M.; Varada, Jan; Marrett, Sean; Thomas, Adam; et al. (2018): MRIQC WebAPI - Database dump. Figshare
doi:10.6084/m9.figshare.7097879