bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MRIQC: Predicting Quality in Manual MRI Assessment Protocols Using No-Reference Image Quality Measures.

Oscar Esteban[1*], Daniel Birman[1], Marie Schaer[2], Oluwasanmi O. Koyejo[3], Russell A. Poldrack[1] and Krzysztof J. Gorgolewski[1]

1 Department of Psychology, Stanford University, Stanford, California, USA
2 Department of Psychiatry, University of Geneva School of Medicine, Geneva, Switzerland
3 Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, USA

* phd@oscaresteban.es

Abstract

Quality control of MR images is essential for excluding problematic acquisitions and avoiding bias in subsequent image processing and analysis. However, the visual inspection of individual images is time-consuming and limited by both intra- and inter-rater variance. The difficulty of visual inspection scales with study size and with the heterogeneity of multi-site data. Here, we describe a tool for the automated assessment of T1-weighted MR images of the brain – *MRIQC*. MRIQC calculates a set of quality measures from each image and uses them as features in a binary (include/exclude) classifier. The classifier was designed to ensure generalization to new samples acquired in different centers and using different scanning parameters from our training dataset. To achieve that goal, the classifier was trained on the Autism Brain Imaging Data Exchange (ABIDE) dataset (N=1102), acquired at 17 locations with heterogeneous scanning parameters. We selected random forests from a set of models and pre-processing options using nested cross-validation on the ABIDE dataset. We report a performance of ∼89% accuracy of the best model evaluated with nested cross-validation. The best performing classifier was then evaluated on a held-out (unseen) dataset, unrelated to ABIDE and labeled by a different expert, yielding ∼73% accuracy. The MRIQC software package and the trained classifier are released as an open source project, so that individual researchers and large consortia can readily curate their data regardless the size of their databases. Robust QC is crucial to identify early structured imaging artifacts in ongoing acquisition efforts, and helps detect individual substandard images that may bias downstream analyses.

Introduction

Image analysis can lead to erroneous conclusions when the original data are of low quality. MRI images are unlikely to be artifact-free, and assessing the quality of images produced by MR scanning systems has long been a challenging issue [1]. Traditionally, all images in the sample under analysis are visually inspected by one or more experts, and those showing an insufficient level of quality are excluded (some examples are given in Fig 1A). Visual assessment is time consuming and prone to variability due to inter-rater differences (see Fig 1B), as well as intra-rater differences arising from factors

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

such as practice or fatigue. An additional concern is that some artifacts evade human detection entirely [2] for example those due to improper choice of acquisition parameters. Even though magnetic resonance (MR) systems undergo periodic inspections and service, some machine-related artifacts persist unnoticed due to lenient

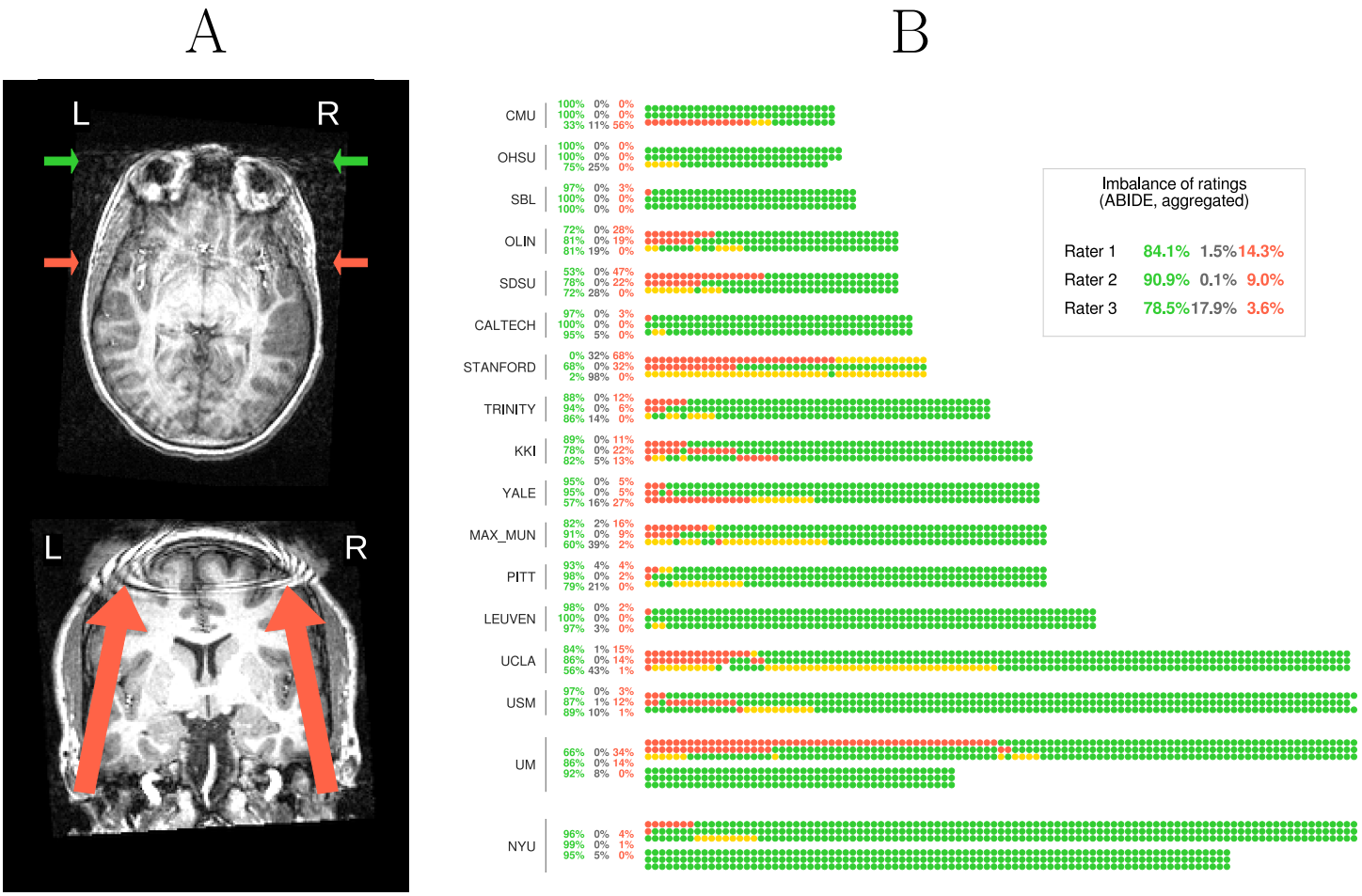

Figure 1. Visual assessment of MR scans and the quality control of the ABIDE dataset.
A. Two images with prominent artifacts from the ABIDE database are presented. An example scan (top) is shown with severe motion artifacts. The arrows point to signal spillover through the phase-encoding axis (right-to-left –RL–) due to eye movements (green) and vessel pulsations (red). A second example scan (bottom) shows severe coil artifacts. This figure caption is extended in Figure SI1.
B. Info-graphic of the visual assessment of the T1-weighted (T1w) MR images of the ABIDE dataset performed by three different experts, split by scanning site. Each scanning site has one stripe with three rows of colored circles, except sites with large samples where the ratings are wrapped in two stripes. Each circle represents the rating of one image by one of the experts, with the color encoding the quality label (green is "accept", yellow is "doubtful", red is "exclude" and white denotes missing ratings). Each row is a different expert, thus the inter-rater consistency can be checked column-wise. A perfect agreement occurs when the three circles of a column show the same color (for example, the first image of the "CALTECH" scanning site). Some images yielded no agreement across raters (e.g. the second participant in the "PITT" sample). Next to each site label, the spread of ratings is reported. Rows are the three raters, and columns the ratings. First (in green color) for "accept", second (gray) for "doubtful", and third (red) is "exclude". The aggregated (all sites of ABIDE) rates are presented in the top-right box.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

vendor quality checks, and drift from the system calibration settings. In our experience, automated Quality Control (QC) protocols help detect these issues early in the processing stream. The current trend of neuroimaging towards acquiring very large samples across multiple scanning sites [3–5] introduces additional concerns. These large scale imaging efforts render the visual inspection of every image infeasible and add the possibility of between-site variability. Therefore, there is a need for fully-automated, robust, and minimally biased QC protocols. These properties are difficult to achieve for three reasons: 1) the absence of a gold standard impedes the definition of sensitive quality metrics; 2) human experts introduce biases with their visual assessment; and 3) cross-study and inter-site acquisition differences also introduce uncharacterized variability.

Machine-specific artifacts have been traditionally tracked down using phantoms [6] in a quantitative manner. However, many forms of image degradation are participant-specific or arise from practical settings (see Fig 1, panel A). Woodard and Carley-Spencer [7] conducted one of the earliest evaluations of automated quality assessment on a large dataset of 1001 T1w images from 143 participants. They defined a set of 239 *no-reference*[1] image-quality metrics (IQMs). The IQMs belonged to two families depending on whether they were derived from Natural Scene Statistics or quality indices defined by the JPEG consortium. The IQMs were calculated on image pairs with and without several synthetic distortions. In an analysis of variance, the IQM from both families reliably discriminated among undistorted images, noisy images, and images distorted by intensity non-uniformity (INU). Mortamet et al. [8] proposed two quality indices focused on detecting artifacts in the air region surrounding the head, and analyzing the goodness-of-fit of a model for the background noise in that air area. One principle underlying their proposal is that most of the artifact signal propagates over the image and into the background. They applied these two IQMs in 749 T1w scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Different cutoff thresholds were defined for the two IQMs and compared to a binary (high/low quality) classification performed by a human rater, concluding that more specific research was required to determine these thresholds and generalize them to different datasets. They achieved an 85% accuracy in an intra-site validation approach. However, many potential sources of uncontrolled variability exist between studies and sites, including magnetic resonance imaging (MRI) protocols (scanner manufacturer, MR sequence parameters, etc.), scanning settings, participant instructions, inclusion criteria, etc. For these reasons, the thresholds they proposed on their IQMs are unlikely to generalize beyond the ADNI database. Recently, Pizarro et al. [9] proposed the use of a support-vector machine classifier (SVC) trained on 1457 structural MRI images acquired in one site with constant scanning parameters. They proposed three volumetric features and three features targeting particular artifacts. The volumetric features were the normalized histogram, the tissue-wise histogram and the ratio of the modes of gray matter (GM) and white matter (WM). The artifacts addressed were the eye motion spillover in the anterior-to-posterior phase-encoding direction, the head-motion spillover over the nasio-cerebellum axis (which they call *ringing artifact*) and the so-called wrap-around (which they refer to as *aliasing artifact*). They reported a prediction accuracy around 80%, assessed using 10-fold cross-validation. Some other recent efforts to develop IQMs appropriate for MRI include the Quality Assessment Protocol[2] (QAP) under the preprocessed-connectomes project (PCP), and the UK Biobank [10].

The hypothesis behind this study is that we can predict the quality ratings of an expert on previously unseen datasets (with dataset-specific scanning parameters) in a

[1] A measure is called "no-reference" when no ground-truth of the same image without degradation is available.

[2] Available online: http://preprocessed-connectomes-project.org/quality-assessment-protocol/.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

supervised learning approach that uses features derived from a broad selection of IQMs. To demonstrate that the trained classifier correctly predicts the quality of new data, we used two unrelated databases to configure the training and held-out (test) datasets [11]. We first select the best performing model on the training dataset using a grid strategy in a nested cross-validation setup. We use the ABIDE database [4] for the training set because data are acquired in 17 different scanning sites with varying acquisition parameters (Table 1). These data show great variability in terms of imaging settings and parameters, what represents the heterogeneity of real world data. The best performing classifier is then trained in the full ABIDE dataset, and tested in the held-out dataset [12] to assess whether the performance on unseen data falls within the range predicted by the nested cross-validation.

The contributions of this work are summarized as follows. First, we release a software tool called MRIQC (described in The MRIQC tool) to extract of a number of IQMs (Extracting the Image Quality Metrics) that characterize each input image. Second, MRIQC includes a visual reporting system (described in the Visual reports section) to ease the manual investigation of potential quality issues. These visual reports allow researchers to quickly evaluate the cases flagged by the MRIQC classifier or visually identify potential images to be flagged by looking at the group distributions of IQMs. Finally, we report the results from a pre-registered analysis of this study (`https://osf.io/haf97/`) on the feasibility of automatic quality control labeling (sections Supervised classification and Results).

Materials and Methods

Training and test datasets

A total of 1375 T1w scans are used as training (1102 from ABIDE) and test (273 from ds030) samples. These databases were intentionally selected for their heterogeneity to match the purpose of the study. A brief summary illustrating the diversity of acquisition parameters is presented in Table 1, and a full-detail table in Table SI1.

Labeling protocol The labeling process is aided by surface reconstruction, using the so-called *white* (WM-GM interface) and the *pial* (delineating the outer interface of the cortex) surfaces as visual cues for the rater. We utilize *FreeSurfer* [13] to reconstruct the surfaces. *FreeSurfer* has been recently proposed as a visual aid tool to assess T1w images [14]. For run-time considerations, and to avoid circular evaluations of *FreeSurfer*, this tool is not used in the MRIQC workflow (see The MRIQC tool section).

The following protocol was used for the manual assessment of T1w images: 1) The 3D cortical surfaces were reconstructed using *FreeSurfer 5.3.0*. 2) An animated GIF (graphics interchange format) file was generated from the coronal slices of the 3D volume, including the projection of the 3D cortical surfaces in each slice[3]. Each animation had a duration of around 20s. 3) A trained expert inspected the animation several times (generally, three times), and assigned a quality level ("exclude"/"doubtful"/"accept").

During the visualization, the rater assessed the overall quality of the image. The *white* and *pial* contours were used as evaluation surrogates, given that "exclude" images usually exhibit imperfections and inaccuracies on these surfaces. When the expert found general quality issues or the reconstructed surfaces revealed more specific artifacts, the "exclude" label was assigned and the rater noted a brief description, for example: "low

[3]We distribute with MRIQC the script `fs2gif` which produces such animations. The animations used to evaluate the ds030 dataset are found here `https://drive.google.com/drive/u/1/folders/0BxI12kyv2olZTDhiUVVMc2FyRDg`.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Summary table of the train and test datasets. The ABIDE dataset is publicly available[a], and contains images acquired at 17 sites, with a diverse set of acquisition settings and parameters. This heterogeneity makes it a good candidate to train machine learning models that can generalize well to novel samples from other sites We selected ds030 [12] from OpenfMRI[b] as held-out dataset to evaluate the performance on data unrelated to the training set. A table summarizing the heterogeneity of parameters within the ABIDE dataset and also ds030 is provided as supplemental material (Table SI1).

Dataset	Site ID	Scanner vendor & model TR/TE/TI [sec], FA [deg], PE dir.	Size[c] [voxels]	Resolution[c] [mm]
ABIDE N=1102	CALTEC	Siemens Magnetom TrioTim, $1.59/2.73{\cdot}10^{-3}/0.8$, 10, AP	176±80×256±32×256±32	1.00×1.00±0.03×1.00±0.03
	CMU	Siemens Magnetom Verio, $1.87/2.48{\cdot}10^{-3}/1.1$, 8, AP	176±15×256±62×256±62	1.00×1.00×1.00
	KKI	Philips Achieva 3T, $8{\cdot}10^{-3}/3.70{\cdot}10^{-3}/0.8$, 8, *NA*	256×200±30×256±30	1.00×1.00×1.00
	LEUVEN	Philips Intera 3T, $9.60{\cdot}10^{-3}/4.60{\cdot}10^{-3}/0.9$, 8, RL	256×182×256	0.98×1.20×0.98
	MAX_MUN	Siemens Magnetom Verio, $1.8/3.06{\cdot}10^{-3}/0.9$, 9, AP	160±16×240±16×256±16	1.00×1.00±0.02×1.00±0.02
	NYU	Siemens Magnetom Allegra, $2.53/3.25{\cdot}10^{-3}/1.1$, 7, AP	128×256×256	1.33×1.00×1.00
	OHSU	Siemens Magnetom TrioTim, $2.3/3.58{\cdot}10^{-3}/0.9$, 10, AP	160×239±1×200±1	1.10×1.00×1.00
	OLIN	Siemens Magnetom Allegra, $2.5/2.74{\cdot}10^{-3}/0.9$, 8, RL	208±32×256×176	1.00×1.00×1.00
	PITT	Siemens Magnetom Allegra, $2.1/3.93{\cdot}10^{-3}/1.0$, 7, AP	176×256×256	1.05×1.05×1.05
	SBL	Philips Intera 3T, $9{\cdot}10^{-3}/3.5{\cdot}10^{-3}/NA$, 7, *NA*	256×256×170	1.00×1.00×1.00
	SDSU	General Electric Discovery MR750 3T, $11.1{\cdot}10^{-3}/4.30{\cdot}10^{-3}/0.6$, 8, *NA*	172×256×256	1.00×1.00×1.00
	STANFORD	General Electric Signa 3T, $8.4{\cdot}10^{-3}/1.80{\cdot}10^{-3}/NA$, 15, *NA*,	256×132×256	0.86×1.50×0.86
	TRINITY	Philips Achieva 3T, $8.5{\cdot}10^{-3}/3.90{\cdot}10^{-3}/1.0$, 8, AP	160×256±32×256±32	1.00×1.00±0.07×1.00±0.07
	UCLA	Siemens Magnetom TrioTim, $2.3/2.84{\cdot}10^{-3}/0.85$, 9, AP	160±16×240±26×256±26	1.20±0.20×1.00±0.04×1.00±0.04
	UM	General Electric Signa 3T, $NA/1.80{\cdot}10^{-3}/NA$, 15, AP	256±154×256×124	1.02±0.38×1.02±0.16×1.20±0.16
	USM	Siemens Magnetom Allegra, $2.1/3.93{\cdot}10^{-3}/1.0$, 7, AP	160±96×480±224×512±224	1.20±0.20×0.50±0.50×0.50±0.50
	YALE	Siemens Magnetom TrioTim, $1.23/1.73{\cdot}10^{-3}/0.6$, 9, AP	160±96×256×256	1.00×1.00×1.00
DS030 N=273	BMAP STAGLIN	Siemens Magnetom TrioTim, $2.53/3.31{\cdot}10^{-3}/1.1$, 7, RL	176×256×256	1.00×1.00×1.00

[a] http://fcon_1000.projects.nitrc.org/indi/abide/. [b] https://openfmri.org/dataset/ds000030/. [c] Sizes and resolutions are reported as follows: median value along each dimension ± the most extreme value from the median (either above or below).

signal-to-nose ratio (SNR)", "poor image contrast", "ringing artifacts", "head motion", etc.). The images in ds030 were randomized before rating.

Software instruments and calculation of the IQMs

The MRIQC tool MRIQC is an open-source project, developed under the following software engineering principles. 1) *Modularity and integrability*: MRIQC implements a `nipype` [15] workflow (see Fig 2) to integrate modular sub-workflows that rely upon third party software toolboxes such as *FSL* [16], *ANTs* [17] and *AFNI* [18]. 2) *Minimal preprocessing*: the workflow described before should be as minimal as possible to estimate the IQMs on the original data or their minimally processed derivatives. 3) *Interoperability and standards*: MRIQC follows the the brain imaging data structure (BIDS, [19]), and it adopts the BIDS-App [20] standard. An example of the ease of

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

running MRIQC is presented in Listing SI1. 4) *Reliability and robustness*: the software undergoes frequent vetting sprints by testing its robustness against data variability (acquisition parameters, physiological differences, etc.) using images from the OpenfMRI resource. Reliability is checked and maintained with the use of a continuous integration service.

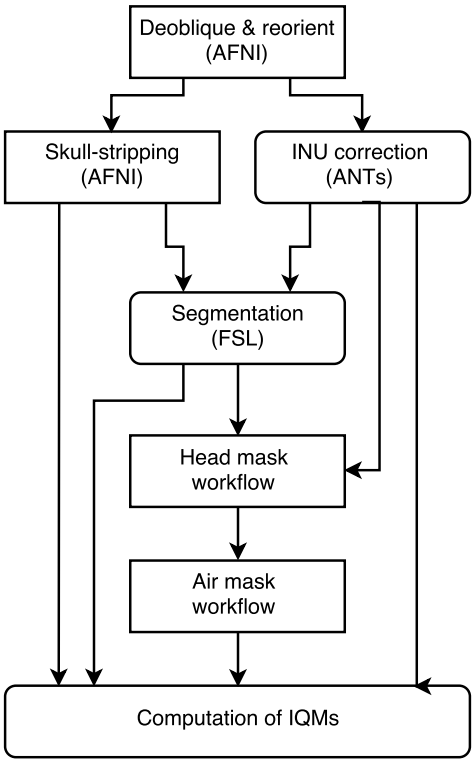

Figure 2. MRIQC's processing data flow. Images undergo an optimized processing pipeline to: 1) realign images in a conformed space using AFNI realign; 2) INU estimation using N4ITK; 3) skull-stripping using AFNI 3dSkullStrip; 4) brain tissue segmentation using FSL-FAST; 5) computation of an air/tissue mask using the magnitude of the gradient image [8]; 6) mapping of an exclusion mask defined in MNI space into the subject using an affine registration scheme with ANTs; 7) computation of an air mask excluding the region below the plane crossing the nasio-cerebellum axis; 8) computation of artifactual regions [8]; and 9) computation of a surrounding air-mask without artifacts; 10) projection of all the computed masks and segmentations to the original (native) space of the image volume

Extracting the Image Quality Metrics The final steps of the MRIQC's workflow compute the different IQMs, and a summary JSON file per subject is generated. The IQMs can be grouped in four broad categories (see Table 2), providing a vector of 56 features per anatomical image. Some measures characterize the impact of noise and/or evaluate the fitness of a noise model. A second family of measures use information theory and prescribed masks to evaluate the spatial distribution of information. A third family of measures look for the presence and impact of particular artifacts. Specifically, the INU artifact, and the signal leakage due to rapid motion (e.g. eyes motion or blood vessel pulsation) are identified. Finally, some measures that do not fit within the previous categories characterize the statistical properties of tissue distributions, volume overlap of tissues with respect to the volumes projected from MNI space, the sharpness/blurriness of the images, etc. The ABIDE and ds030 datasets were processed with *MRIQC–v.0.9.0-rc2* using the *Lonestar5* supercomputer at the Texas Advanced Computing Center, University of Texas, TX, USA.

Visual reports. In order to ease the screening process of individual images, MRIQC generates individual reports with mosaic views of a number of cutting planes and supporting information (for example, segmentation contours). The most straightforward use-case is the visualization of those images flagged as low-quality by the classifier.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2. Summary table of IQMs. The 14 IQMs spawn a vector of 56 features per anatomical image on which the classifier is learned and tested.

Measures based on noise measurements	
CJV	The coefficient of joint variation of GM and WM was proposed as objective function by Ganzetti et al. [21] 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 [22] 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 signal-to-nose ratio calculation proposed by Dietrich et al. [23], 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 [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.
Measures based on information theory	
EFC	The entropy-focus criterion [24] 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 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.
Measures targeting specific artifacts	
INU	MRIQC measures the location and spread of the bias field extracted estimated by the intensity non-uniformity correction. The smaller spreads located around 1.0 are better.
QI_1	The first quality index of [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 measures	
FWHM	The full-width half-maximum is an estimation of the blurriness of the image using AFNI's `3dFWHMx`. Smaller is better.
ICVs	Estimation of the intracranial volume 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 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].

After the extraction of IQMs in all the images of our sample, a group report is generated (Fig 3). The group report shows a scatter plot for each of the IQMs, so it is particularly easy to identify the cases that are outliers for each metric. The plots are interactive, such that clicking on any particular sample opens the corresponding individual report of that case. Examples of group and individual reports for the ABIDE dataset are available online at mriqc.org.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

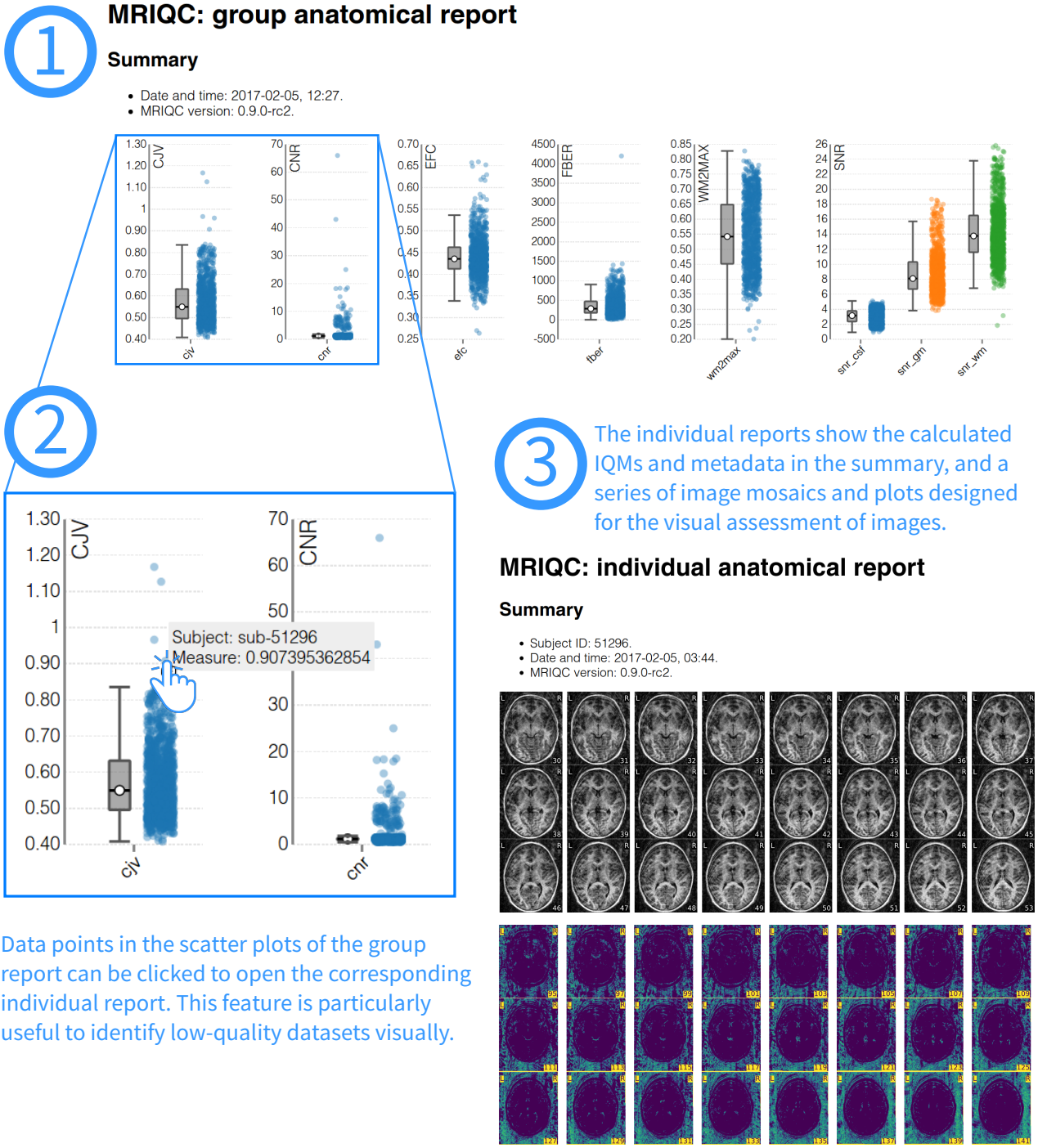

Figure 3. Visual reports. MRIQC generates one individual report per subject in the input folder and one group report including all subjects. To visually assess MRI samples, the first step (1) is opening the group report. This report shows boxplots and strip-plots for each of the IQMs. Looking at the distribution, it is possible to find images that potentially show low-quality as they are generally reflected as outliers in one or more strip-plot. For instance, in (2) hovering a suspicious sample within the coefficient of joint variation (CJV) plot, the subject identifier is presented ("sub-51296"). Clicking on that sample will open the individual report for that specific subject (3). This particular example of individual report is available online at `https://web.stanford.edu/group/poldracklab/mriqc/reports/sub-51296_T1w.html`.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supervised classification

Our supervised learning approach to predicting the binary ratings of a human expert is structured in two steps. First, we perform a preliminary model selection and evaluation using repeated (x1000) and nested cross-validation, on the ABIDE dataset (see Step 1: Tested models and selection). Then, a second optimization in a refined grid of hyper-parameters for the model selected previously is performed with a single-loop cross-validation on the ABIDE dataset. The best performing model of this second cross-validation step is evaluated using the held-out dataset (see Step 2: Validation on the held-out dataset). The cross-validation workflows are built upon `scikit-learn` [26] and run using the *Stampede* supercomputer at the Texas Advanced Computing Center, University of Texas, TX, USA.

Step 1: Tested models and selection

Based on the number of features (56) and training data available (∼1100 data points), we compare two families of classifiers: SVCs and random forests classifiers (RFCs). Given the diversity of scanning sites, in the model selection loop we also investigate the need for normalizing (*zscoring*) features. In the following, models including a preliminary *zscoring* will show the suffix "-*zs*" while those using the original features without such transformation are noted with the suffix "-*nzs*".

The support-vector machine classifier (SVC) A support-vector machine [27] finds a hyperplane in the high-dimensional space of the features that robustly classifies the data. The SVC then uses the hyperplane to decide the class that is assigned to new samples in the space of features. Two hyper-parameters define the support-vector machine algorithm: a kernel function that defines the similarity between data points to ultimately compute a distance to the hyperplane, and a regularization weight C. In particular, we analyzed here the linear SVC implementation (as of now, "SVC-lin") and the one based on radial basis functions (denoted by "SVC-rbf"). During model selection, we evaluated the regularization weight C and the γ parameter (kernel width) of the SVC-rbf.

The random forests classifier (RFC) Random forests [28] are an nonparametric ensemble learning method that builds multiple decision trees. Then, a RFC assigns to each new sample the mode of the predicted classes of all decision trees in the ensemble. In this case, random forests are driven by a larger number of hyper-parameters. Particularly, in this work we analyze the maximum tree-depth, the minimum number of samples per split and the total number of decision trees.

Objective function The performance of each given model and parameter selection can be quantified with different metrics. Given the imbalance of positive and negative cases –with lower prevalence of "reject" samples–, we select the area under the curve (AUC) of the receiver-operator characteristic as objective score. We also report the classification accuracy as an additional performance measure.

Cross-validation and nested cross-validation Cross-validation is a model selection and validation technique robust to inhomogeneities [29]. We use nested cross-validation, which divides the process in two validation loops: an inner loop for selecting the best model and hyper-parameters, and an outer loop for evaluation. In cross-validation, the data are split into a number of folds, each containing a training and a test set. For each fold, the classifier is trained on the first set and evaluated on the latter. When cross-validation is nested, the training set is split again into folds

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

within the inner loop, and training/evaluation are performed to optimize the model parameters. Only the best performing model of the inner loop is then cross-validated in the outer loop. In order to increase the robustness against model variability, we repeat the nested cross-validation procedure 1000 times.

Data split scheme Since we wanted to estimate the performance in datasets acquired at sites and with parameters different from those in the ABIDE dataset, we selected a *leave-one-site-out (LoSo)* partition strategy for the outer loop of cross-validation. The LoSo split leaves a whole site as a test set at each cross-validation iteration. Therefore, no knowledge of the testing set is leaked into the training set (the remaining $N-1$ sites). For the inner loop (model selection) we compared the performance of a stratified 10-fold and LoSo over the remaining 16 sites (one site is held out by the outer loop). All the possible combinations of models and their hyper-parameters (over 5000) are evaluated repeatedly (1000 times) in a grid search for the best average AUC score in the inner cross-validation loop.

Feature ranking One tool to improve the interpretability of the RFC is the calculation of feature rankings [28] by means of variable importance or Gini importance. Since we use `scikit-learn`, the implementation uses Gini importance, defined for a single tree as the total decrease in node impurity weighted by the probability of reaching that node. We finally report the median feature importance over all trees of the ensemble.

Step 2: Validation on the held-out dataset

In the second step, we cross-validated the model selected in step 1, optimizing a grid search refined to the selection of parameters done before. For this second cross-validation, we use the LoSo split strategy given the results obtained in the previous step. The best performing model is then trained on the full ABIDE dataset and the resulting classifier is used in the prediction of quality ratings of the held-out dataset (ds030).

Results

All images included in the selected datasets were processed with MRIQC. After extraction of the IQMs from the ABIDE, a total of 1102 images had both quality ratings and quality features (ten T1w images of the ABIDE are missing in the database). In the case of ds030, 265 images had the necessary quality ratings and features (eight images were not rated and/or failed during feature extraction).

Model selection

The results of the step 1 (nested cross-validation) are summarized in Fig 4. The best performing model, regardless of inner loop split strategy, was the random forests classifier without *zscoring* (RFC-nzs). The RFC-nzs using LoSo in the inner loop yielded the following averaged scores off all repeated outer loops: AUC=0.862 (σ=$\pm$0.121) and accuracy of 89.4% (σ=$\pm$9.95%). The corresponding averaged scores for the 10-fold strategy were: AUC=0.848 (σ=$\pm$0.135) and accuracy of 88.6% (σ=$\pm$11.5%). These results indicated that there is no practical difference between the two split strategies as regards model selection through cross-validation on this dataset. Therefore, since the averaged scores using LoSo cross-validation in the inner loop were slightly higher, it was selected as split strategy for the cross-validation in step 2. Note that the split strategy is not a model feature, and thus this decision can

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

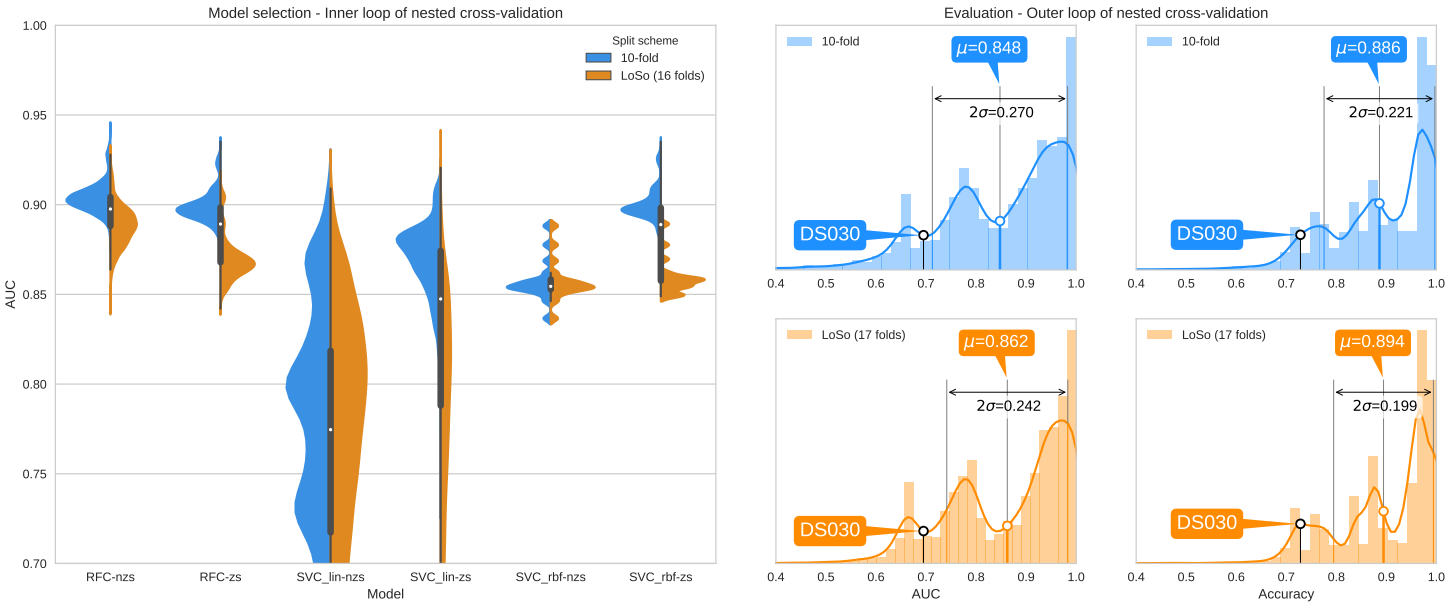

Figure 4. Results of the step 1, nested cross-validation. (Left) Model selection – inner loop. All the possible combinations of model, *zscoring* and hyper-parameters were evaluated. The violinplot shows the distribution of scores of the best performing hyper-parameters per model and preprocessing combination. The scores obtained using the stratified 10-fold split are presented in blue. In orange, the results corresponding to LoSo. In general, the 10-fold splitting was more optimistic for all models, whereas the LoSo scores are closer to results obtained in the outer (evaluation) loop. In all iteration loops, regardless of split strategy and cross-validation repetition, the RFC-nzs achieved the best score, with varying parameters. As expected, *zscoring* the features was necessary for both SVC_lin and SVC_rbf to exhibit acceptable performances, but always below that of the RFC-nzs. **(Right) Evaluation – outer loop**. On the right hand panel, colors represent again the split strategy used in the inner loop. With colored markers, the average cross-validated score is annotated in a box with the μ symbol. Below, the spread of the distribution is noted. Please note that, since the scores are bounded above 1.0, the values of the standard deviation σ are probably underestimated. The distributions of nested cross-validated scores for both AUC and accuracy were rather independent from the split strategy used in the inner loop. The results for both metrics obtained in the evaluation of the held-out dataset (ds030) are represented in the corresponding distribution of nested cross-validation scores, showing that the performance on unseen data falls very close to one standard deviation below the average score. In this case, the average cross-validated score was higher for LoSo (AUC=0.862/accuracy≈89.4%) as compared to the 10-fold split (AUC=0.848/accuracy≈88.6%). Also the spread of cross-validated scores is slightly lower for LoSo (AUC=±0.121/accuracy=±9.95% vs. AUC=±0.135/accuracy=±11.5%).

be made based on the results of the outer loop of nested cross-validation, as opposed to the model selection that is done based on the inner cross-validation loops.

The best performing model and parameters selected as the maximum average of the AUC score in the inner loop, across all repetitions of the nested cross-validation was the RFC-nzs, with 50 trees (`n_estimators`), maximum tree depth (`max_depth`) of 20, and a minimum of 2 samples per split (`min_samples_split`). However, the cross-validation was very variable in the selection of hyper-parameters, indicating that there was little difference in performance for all the points in the hyper-parameters grid.

Evaluation on held-out data

In the second cross-validation step, only the previously selected RFC-nzs model was optimized, in a refined grid centered around the best performing parameters of step 1

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

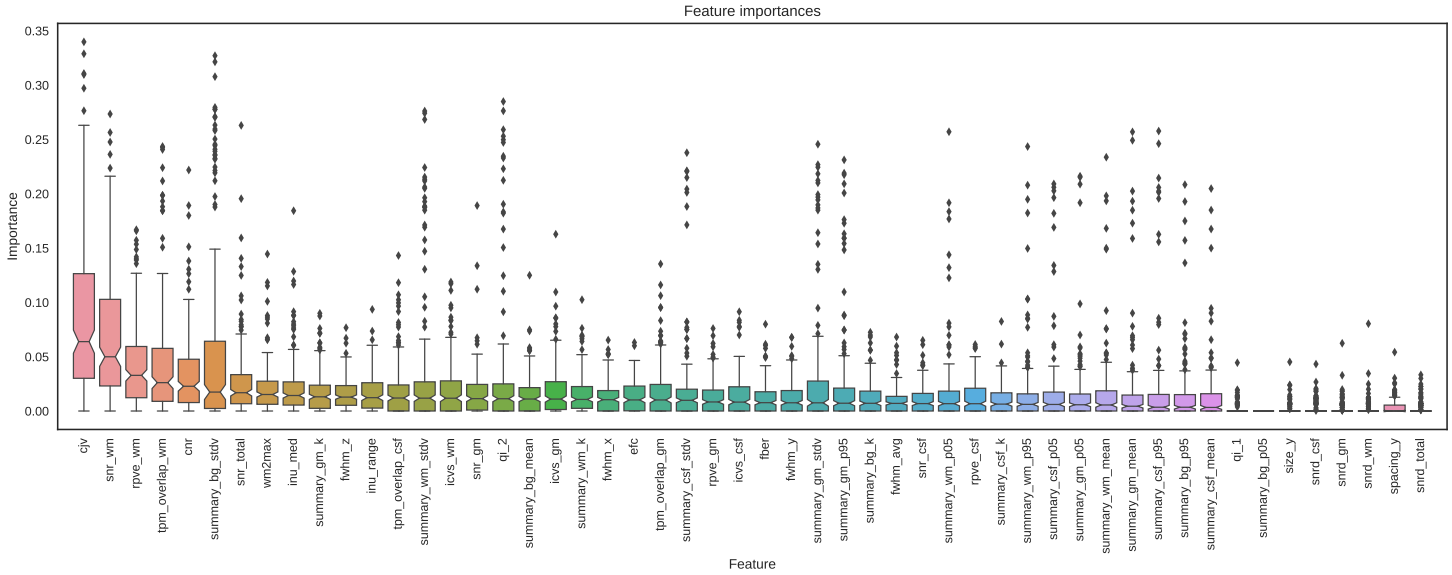

Figure 5. Feature importances in the final classifier. The features used by the best-performing RFC-nzs classifier used to evaluate the held-out dataset are presented. For each feature the boxplot represents the distribution of feature importances within the trees in the ensemble. The features are ordered from highest median importance (the CJV) to lowest (average SNR computed as in [23], `snrd_total`).

(`n_estimators`=50, `max_depth`=20, `min_samples_split`=2). The AUC on the evaluation set was 0.695 and the accuracy 72.83%. We also analyzed the relevance of each feature in the overall forest decision (Fig 5). The most relevant features are the coefficient of joint variation (CJV) and the SNR measured on the WM tissue mask.

Discussion

We propose a quantitative approach to quality control T1w MRI of the brain, enabling the automatic identification of sub-standard acquisitions. Quality control protocols are implemented to exclude faulty datasets that can bias the final results. Human brain images can be degraded by various sources of artifacts, related to the scanning device and settings or due to the participants themselves. Machine-derived artifacts are efficiently mitigated in a quantitative manner with calibration. However, due to the lack of reliable quality quantification tools, subject-specific artifacts and drifts from the service settings are assessed visually. The visual inspection of every MRI acquisition of the brain is a time-consuming and bias-prone task that would be ideally replaced by decision algorithms. Automating the QC process is particularly necessary for ongoing studies such as the UK Biobank that will collect data from tens of thousands of individuals.

Previous efforts [7, 30] in the quantification of image quality for their assessment included the definition of image-quality metrics (IQMs). However, the approach was unfeasible for a total automation due to the limited sensitivity of the available IQMs to the most prevalent artifacts. Subsequent efforts [8] were focused on specific samples, setting generalization to new datasets as a future line of their work. Pizarro et al. [9] recently presented a similar approach to quality control images. They obtain a cross-validated accuracy of ∼80% for their support-vector machine classifier (SVC), in a single-site sample with homogeneous acquisition parameters.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In this work, we train a random forests classifier and evaluate its performance to predict the quality assessment of human raters on completely novel samples. We show that linear SVCs do not perform well on heterogeneous samples with diversity of acquisition parameters, and they always require normalization of features derived from multi-site data. Our results invariably indicated a better performance of a random forests classifier (RFC), with and without normalization of features. Particularly, the best performing model (RFC-nzs, for "not zscored") achieved a ∼89.4% (σ=±9.95%) accuracy. This improved performance over the one reported by Pizarro et al. may also be related to the selection of classification features proposed in this paper. Even though they reported that classification improved with the addition of features addressing certain artifacts, in our feature importance analysis the first IQM addressing a specific artifact was ranked in 9th position. This result suggests that there are complex relationships between the features (in multi-site studies) that may not be captured by SVCs. When tested on unseen data, the RFC-nzs classifier yielded an area under the curve (AUC) score of ∼0.695 and accuracy of 73%. This performance falls within the performance previously evaluated with nested cross-validation. We could not compare these results with [9] since they did not test their resulting classifier on a held-out dataset. The performance drop between the nested cross-validated score (∼89%) and the score obtained on the held-out data (∼73%) may be explained by the interplay of several factors. First, we introduced an unplanned inter-rater bias since the held-out dataset could not be rated by the same expert who rated the ABIDE dataset. This limitation could be reduced by calibrating the ratings of the held-out data having the second expert rate a random subsample of the training dataset. Second, the share of scanning vendors, models and corresponding images in the ABIDE dataset is not uniform. The use of a more uniform training dataset could potentially help generalize better to new datasets.

We used nested cross-validation to select the most predictive classifier, ensuring that the evaluation loop was unbiased using a leave-one-site-out (LoSo) splitting strategy. In this cross-validation scheme, the accuracy is bound below by that measured during the test validation loop. Therefore, the final classifier is ultimately trained using all the available data to push its predictive accuracy above the evaluated performance.

This quantitative assessment of quality is the central piece of the three-fold contribution of this paper. The first outcome of this study is the MRIQC toolbox, a set of open-source tools which compute quality features. Second, MRIQC generates interactive visual reports that allow further interpretation of the decisions made by the classifier. Finally we propose the automated quality control tool described before to generate include/exclude decisions. The MRIQC toolbox is a fork of the Quality Assessment Protocol (QAP). Since MRIQC was started as a standalone project, the implementation of most of the IQMs have been revised, and some are supported with unit tests. As QAP, MRIQC also implements a functional MRI (fMRI) workflow to extract IQMs and generate their corresponding visual reports. Some new IQMs have been added (for instance, the CJV, those features measuring the INU artifacts, or the rPVEs). The group and individual visual reports for structural and functional data are also new contributions to MRIQC with respect to the fork from QAP. The last diverging feature of MRIQC with respect to QAP is the cross-validation work and the release of the trained classifier.

MRIQC is one effort to standardize methodologies that make data-driven and objective QC decisions. Automated QC can provide unbiased exclusion criteria for neuroimaging studies, helping avoid "cherry-picking" of data. A second potential application is the use of automated QC predictions as data descriptors to support the recently born "data papers" track of many journals and public databases like OpenfMRI [31]. The ultimate goal of the proposed classifier is its inclusion in automatic

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

QC workflows, before image processing and analysis. Ideally, minimizing the run time of MRIQC, the extraction and classification process could be streamlined in the acquisition, allowing for the immediate repetition of ruled out scans. Integrating MRIQC in our research workflow allowed us to adjust reconstruction methodologies, tweak the instructions given to the participant during scanning, and minimize the time required to visually assess one image with the visual reports.

Conclusion

We propose MRIQC, a quality control software tool to assess structural MRI of the human brain. MRIQC generates visual reports to speed the screening process, and a set of features which were used to train an automated decision tool. We trained a random forests classifier on the ABIDE dataset (N=1102), acquired at 17 scanning sites with diverse acquisition parameters. We utilized repeated-and-nested cross-validation, with a leave-one-site-out splitting strategy. This avoided hidden feature relationships leaking from the site under test to the training set, ensuring that the evaluated performance was agnostic to site and ultimately represented well the generalization of performance to unseen data. The nested cross-validation evaluation yielded a $\sim$89.4% (σ=$\pm$9.95%) accuracy. We double checked this generalization evaluating the performance of the classifier in a previously unseen dataset (N=265) unrelated to ABIDE. The performance on the held-out dataset was $\sim$73% accuracy. This performance fell within the spread of the cross-validated evaluation. We release MRIQC open-source, along with the best performing classifier. The automatic QC of MRI scans, and the implementation of tools to assist the visual assessment of individual images are two tools in high demand for neuroimaging research.

Author contributions

OE lead the development of MRIQC, implemented the cross-validation workflow, pre-registered the report, drafted the manuscript, run the experiments and interpreted the results. KJG devised the machine learning approach to quality control, coordinated the project, contributed to MRIQC and the cross-validation workflow, pre-registered the report, and interpreted the results. MS rated the ABIDE dataset, helped understanding the problems of inter- and intra- rater variabilities. DB rated the ds030 dataset. OOK contributed in the design of the cross-validation workflow and interpreted the results. RAP devised and coordinated the project, advised in all aspects of MRIQC, the cross-validation workflow and the manuscript design, pre-registered the report and interpreted the results. All the authors have read and edited the manuscript.

Availability of MRIQC and the trained classifier

MRIQC is available under the BSD 3-clause license. Source code is publicly accessible through GitHub (`https://github.com/poldracklab/mriqc`). We provide four different installation options: 1) using the source code downloaded from the GitHub repository; 2) using the PyPi distribution system of Python; 3) using the `poldracklab/mriqc` Docker image; or 4) using BIDS-Apps [20]. For detailed information on installation and the user guide, please access `http://mriqc.rtfd.io`. A distributable version of the classifier is also released, trained on all the available data (including the full-ABIDE and the ds030 datasets).

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgments

This work was supported by the Laura and John Arnold Foundation. The authors want to thank the QAP developers (C. Craddock, S. Giavasis, D. Clark, Z. Shezhad, and J. Pellman) for the initial base of code which MRIQC was forked from, W. Triplett and CA. Moodie for their initial contributions with bugfixes and documentation, and J. Varada for his contributions on th source code. JM. Shine and PG. Bissett reviewed the first draft of this manuscript, and helped debug early versions of MRIQC. S. Bhogawar, J. Durnez, I. Eisenberg and JB. Wexler routinely use and help debug the tool.

References

1. Kaufman L, Kramer DM, Crooks LE, Ortendahl DA. Measuring signal-to-noise ratios in MR imaging. Radiology. 1989;173(1):265–267. doi: 10.1148/radiology.173.1.2781018.

2. 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. doi: 10.1016/S1076-6332(05)80184-9.

3. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, et al. The Human Connectome Project: A data acquisition perspective. NeuroImage. 2012;62(4):2222–2231. doi: 10.1016/j.neuroimage.2012.02.018.

4. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry. 2014;19(6):659–667. doi: 10.1038/mp.2013.78.

5. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nature Neuroscience. 2016;advance online publication. doi: 10.1038/nn.4393.

6. Price RR, Axel L, Morgan T, Newman R, Perman W, Schneiders N, et al. Quality assurance methods and phantoms for magnetic resonance imaging: Report of AAPM nuclear magnetic resonance Task Group No. 1. Medical Physics. 1990;17(2):287–295. doi: 10.1118/1.596566.

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

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. doi: 10.1002/mrm.21992.

9. 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. doi: 10.3389/fninf.2016.00052.

10. Alfaro-Almagro F, Jenkinson M, Bangerter N, Andersson J, Griffanti L, Douaud G, et al. UK Biobank Brain Imaging: Automated Processing Pipeline and Quality Control for 100,000 subjects. In: Organization for Human Brain

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Mapping. Geneve, Switzerland; 2016. p. 1877. Available from: https://ww5.aievolution.com/hbm1601/index.cfm?do=abs.viewAbs&abs=3664.

11. Ripley BD. Pattern recognition and neural networks. 7th ed. United Kingdom: Cambridge University Press; 2007.

12. Poldrack RA, Congdon E, Triplett W, Gorgolewski KJ, Karlsgodt KH, Mumford JA, et al. A phenome-wide examination of neural and cognitive function. Scientific Data. 2016;3:160110. doi: 10.1038/sdata.2016.110.

13. Fischl B. FreeSurfer. NeuroImage. 2012;62(2):774–781. doi: 10.1016/j.neuroimage.2012.01.021.

14. 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. doi: 10.3389/fnins.2016.00558.

15. Gorgolewski KJ, Esteban O, Burns C, Ziegler E, Pinsard B, Madison C, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. Zenodo [Software]. 2016;doi: 10.5281/zenodo.50186.

16. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. NeuroImage. 2012;62(2):782–790. doi: 10.1016/j.neuroimage.2011.09.015.

17. Avants B, Duda J, Song G, Das S, Pluta J, Tustison N. ANTs: Advanced Normalization Tools [software]; 2013. Available from: http://www.picsl.upenn.edu/ANTS/.

18. Cox RW, Hyde JS. Software tools for analysis and visualization of fMRI data. NMR in Biomedicine. 1997;10(4-5):171–178. doi: 10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L.

19. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data. 2016;3:160044. doi: 10.1038/sdata.2016.44.

20. Gorgolewski KJ, Alfaro-Almagro F, Auer T, Bellec P, Capota M, Chakravarty M, et al. BIDS Apps: Improving ease of use, accessibility and reproducibility of neuroimaging data analysis methods. bioRxiv. 2016; p. 079145. doi: 10.1101/079145.

21. 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.

22. 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.

23. 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.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24. 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.

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. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

27. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273–297. doi: 10.1007/BF00994018.

28. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32. doi: 10.1023/A:1010933404324.

29. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, et al. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. NeuroImage. 2017;147:736–745. doi: 10.1016/j.neuroimage.2016.10.045.

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

31. Poldrack RA, Barch DM, Mitchell J, Wager T, Wagner AD, Devlin JT, et al. Toward open sharing of task-based fMRI data: the OpenfMRI project. Frontiers in Neuroinformatics. 2013;7:12. doi: 10.3389/fninf.2013.00012.

bioRxiv preprint doi: https://doi.org/10.1101/111294; this version posted February 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supporting Information

Table SI1 Image acquisition parameters. A table containing all the acquisition parameters is maintained in GitHub: https://github.com/oesteban/mriqc/blob/c9bdfa863ca47894d5cdcb605071a5088840afcc/mriqc/data/csv/scan_parameters.tsv.

Listing SI1 Running MRIQC. The BIDS standard makes MRIQC compatible with almost any input dataset without need for custom settings. Since all the metadata associated to the dataset are found in `bids-data/`, the following example would nicely run without further settings. The second positional argument, `out/` indicates where the outputs will be written, and finally, the `participant` keyword instructs MRIQC to run the first level analysis as specified in BIDS Apps.

```
mriqc bids-data/ out/ participant
mriqc bids-data/ out/ participant --participant_label S001 S002
```

Listing SI2 Running MRIQC – Group Level. If the `participant` level was run setting some `--participant_label`, the `group` level is not triggered by default. It can be done manually, pointing the input data folder to the derivatives folder generated with the `participant` level analysis:

```
mriqc out/derivatives/ out/ group
```

Listing SI2 Predicting quality. Although the `group` runlevel will generate a CSV table with the quality label predicted for each sample, it is possible to run the classifier individually:

```
mriqc_clf --load-classifier -X aMRIQC.csv -o mypredictions.csv
```

The default classifier can be replaced by a custom one using:

```
mriqc_clf --load-classifier my_custom_classifier.pklz -X aMRIQC.csv -o
     mypredictions.csv
```

The documentation website contains more detailed information on how to train custom classifiers, or generate refined results from prediction: http://mriqc.readthedocs.io/en/latest/classifier.html.

Figure SI1 Extended caption of Fig 1A. An example scan (top) is shown with severe motion artifacts. The reduced contrast between tissues and the ringing intensity waves in the anterior region of the brain in the presented slice suggest a large head movement occurred during acquisition. The green arrows point to signal spillover due to eye movements through the phase-encoding axis (in this case, right-to-left –RL–). Oftentimes, the RL or LR axes are selected for phase-encoding because the signal leakage from the eyeballs does not overlap with brain tissue, as opposed to selecting anterior-posterior directions. However, the red arrows point to signal spillover caused by vessel pulsations. Given the location of the vessel, in this case signal leakage overlaps brain tissue and affects the quality of this image. The phase-encoding axis has less bandwidth and thus, is more sensitive to movement. For that reason, it is generally selected to have the shortest field of view. A second example scan (bottom) shows severe coil artifacts.