Improving out-of-sample prediction of quality of MRIQC

Oscar Esteban[1], Russell A. Poldrack[1], and Krzysztof J. Gorgolewski[1]

Stanford University, phd@oscaresteban.es

Abstract. MRIQC is a quality control tool that predicts the binary rating (accept/exclude) that human experts would assign to T1-weighted MR images of the human brain. For such prediction, a random forests classifier performs on a vector of image quality metrics (IQMs) extracted from each image. Although MRIQC achieved an out-of-sample accuracy of ∼76%, we concluded that this performance on new, unseen datasets would likely improve after addressing two problems. First, we found that IQMs show "site-effects" since they are highly correlated with the acquisition center and imaging parameters. Second, the high inter-rater variability suggests the presence of annotation errors in the labels of both training and test data sets. Annotation errors may be accentuated by some preprocessing decisions. Here, we confirm the "site-effects" in our IQMs using t-student Stochastic Neighbour Embedding (t-SNE). We also improve by a ∼10% accuracy increment on the out-of-sample prediction of MRIQC by revising a label binarization step in MRIQC. Reliable and automated QC of MRI is in high demand for the increasingly large samples currently being acquired. We show here one iteration to improve the performance of MRIQC on this task, by investigating two challenging problems: site-effects and noise in the labels assigned by human experts.

Keywords: MRIQC · site-effect · labels noise.

1 Introduction

Imaging artifacts have the potential to hamper validity of MRI analyses [1–4]. Moreover, the early detection of subpar images is crucial to identify and avoid propagation of structured degradations derived from the scanning settings, parameters, infrastructure or software [5] in on-going scanning initiatives. Typically, quality control (QC) has been done manually through visual inspection of all individual images within a sample. This approach is both highly unreliable due to intra- and inter- rater variabilities (as we showed in [5]) and prohibitive in large imaging projects. To address these concerns, we presented MRIQC [5], an easy-to-use software that automates the extraction of image quality metrics (IQMs). MRIQC includes a (pre-trained) random-forests classifier to predict the quality of an input image based on its corresponding IQMs. Two experts rated all T1-weighted (T1w) images collected from the ABIDE dataset [6] and OpenNeuro's ds000030 [7, 8] to train and evaluate the classifier. Raters achieved a

Kappa reliability score of 0.51 (moderate) and MRIQC performed with 76% accuracy and a very low sensitivity to the "bad quality" class of 0.28. The ABIDE dataset was used for model selection and training, while the performance was estimated on ds000030 as testing, held-out dataset. We also found that ds000030 contains an idiosyncratic artifact (a ghost of a headset overlapping the temporal lobes) present in many of the images that was "unseen" by the classifier on ABIDE (the training sample). Removing the images with the artifact resulted in an improved accuracy of 87% and sensitivity of 0.46. We concluded that the diversity of the ABIDE dataset, acquired at 17 different sites, was not enough for the classifier to generalize out-of-sample entailing a problem of "batch-effects" for the classifier [9]. We also found that the manual ratings are noisy (i.e. incorrect quality labels are assigned due to human variability), with an impact on the performance of the classifier that we try to understand in this work.

Image quality control methodologies can be roughly divided into two groups, depending on whether their classification is required to work universally or within similar images acquired with similar infrastructure, settings and parameters. Early proposals attempted to solve this "within-sample" outlier detection problem, by defining different features and cut-off thresholds that worked for their image set. Woodard and Carley-Spencer [10] were first on defining and extracting IQMs. They tested their IQMs applying ANOVA on 1001 images from 143 participants, comparing those derived from original images against those derived from images degraded with simulated artifacts (several levels of noise and geometrical distortions). Mortamet et al. [11] 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. 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 on 749 T1w scans from the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset. By defining cut-off thresholds for the two IQMs, they assigned the images high or low quality labels, and compared this classification to a manual assessment. Later efforts to develop IQMs appropriate for MRI include the Quality Assessment Protocol (QAP), and the UK Biobank [12]. Recently, Pizarro et al. [13] proposed the use of a support-vector machine classifier (SVC) trained on 1457 structural MRI images acquired at one same-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. These previous efforts succeeded in showing that automating quality ratings of T1w MRI scans is possible. However, they did not achieve generalization to new datasets acquired at sites unseen by the classifier.

Concerned with generalization, in [5] we proposed MRIQC along with a machine learning framework: i) to ensure un-biased estimation of performance through adequate cross-validation; and ii) which was tested on a held-out dataset to demonstrate out-of-sample performance. MRIQC extends the list of IQMs from the QAP, which was constructed from a careful review of the MRI and medical imaging literature [14]. Here, we investigate hidden structures in the IQMs extracted from our training set (ABIDE) and the impact on prediction accuracy and sensitivity of labels-noise for MRIQC. After curation, the performance of the classifier improved modestly, suggesting that the "site-effects" have a larger influence on performance than the noisy labels. Therefore, we conclude that the problem demands for larger, multi-site, and expert-annotated databases to train automatic decision algorithms, along with new IQMs or features that are more directly related to actual quality properties and artifacts found in images.

2 Methods

2.1 Data

We reuse the MRIQC database, consisting of 1367 pairs of IQM vectors derived from T1w images from both ABIDE and DS00030, and 1467 corresponding manual ratings for all of them, with an overlap of 100 samples rated by both experts (see [5]). These ratings are summarized in Figure 2.

2.2 MRIQC: extraction of image quality metrics and inference

The quality control implemented with MRIQC is a two step process. First, a minimal image processing workflow is used to extract a vector of 64 IQMs from the input dataset (see Figure 1). The nature and implementation of these IQMs are described in the original MRIQC paper. Then, based on the MRIQC database described above, we performed model selection, training and evaluation of a classifier. Model selection through leave-one-site-out validation [5] led to a random forests classifier preceded with three preprocessing steps. First, a feature selection step based on extremely randomized forests to predict the site of origin from the IQMs is used to remove features highly correlated with their site of origin. Then, a second feature selection steps removes features less predictive than a synthetic, random noise-feature. Finally, site-wise scaling to normalize the site-wise standard deviation to 1.0 was also selected.

2.3 Manifold learning approach to investigate the structure of the IQMs

We use t-distributed Stochastic Neighbor Embedding (t-SNE) as implemented in *scikit-learn* [15] to project the samples in our training set (ABIDE dataset) to a bidimensional space easier to interpret, in order to investigate inherent structures on the data. T-SNE translates affinities in the original space represented by Students t-distributions into Gaussian joint probabilities. To do so,

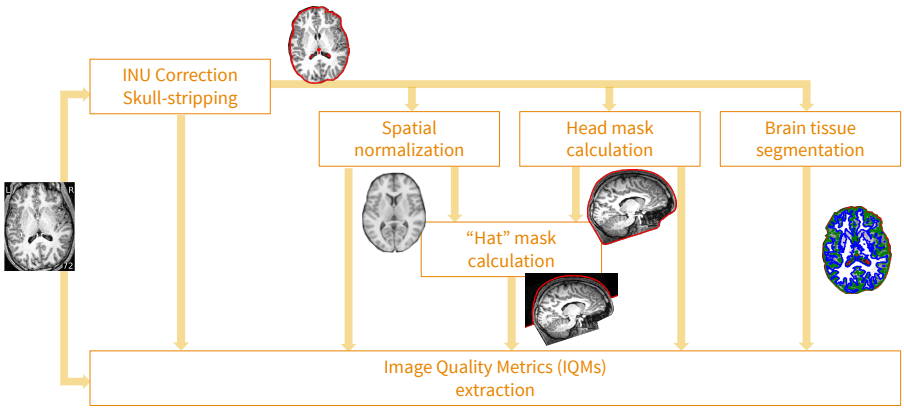

Fig. 1. MRIQC implements a minimal processing workflow that enables the extraction of a vector of 64 IQMs from unprocessed T1-weighted MRIs of the human brain.

the Kullback-Leibler divergence of the joint probabilities in the original space and the embedded space are minimized through gradient descent. We configured t-SNE with random initialization, 5000 maximum iterations, and the default perplexity of 30. Finally, we plot the projection, using colors to represent particular conditions.

2.4 Curating existing quality annotations

In our first proposal of MRIQC [5], the labeling protocol allowed the raters for using three possible annotations: "accept", "doubtful" and "exclude". Since the proposed classifier only predicted quality with a binary label ("accept"/"exclude"), we binarized the labels mapping all "doubtful" cases into the "accept" class. This decision yielded rather imbalanced quality classes ("accept"/"exclude": 63%/37% for ABIDE, 83%/17% for ds000030), and introduced an additional source of quantification errors in our labels as one would expect to have both "accept" and "exclude" cases annotated as "doubtful". Therefore, one additional expert (OE) visually assessed (using MRIQC-generated individual reports) those data points where either rater 1 or 2 assigned the "doubtful" label. Only "accept" and "exclude" ratings were assigned in this new assessment. After the relabeling, the MRIQC classifier was trained again and the out-of-sample performance was estimated on the held-out dataset (ds000030), for which all "doubtful"-rated cases were revised (and assigned either "exclude" or "accept"). We re-estimated the performance of the original MRIQC classifier using the curated quality annotations of ds000030 for a fair comparison to the newly trained classifier.

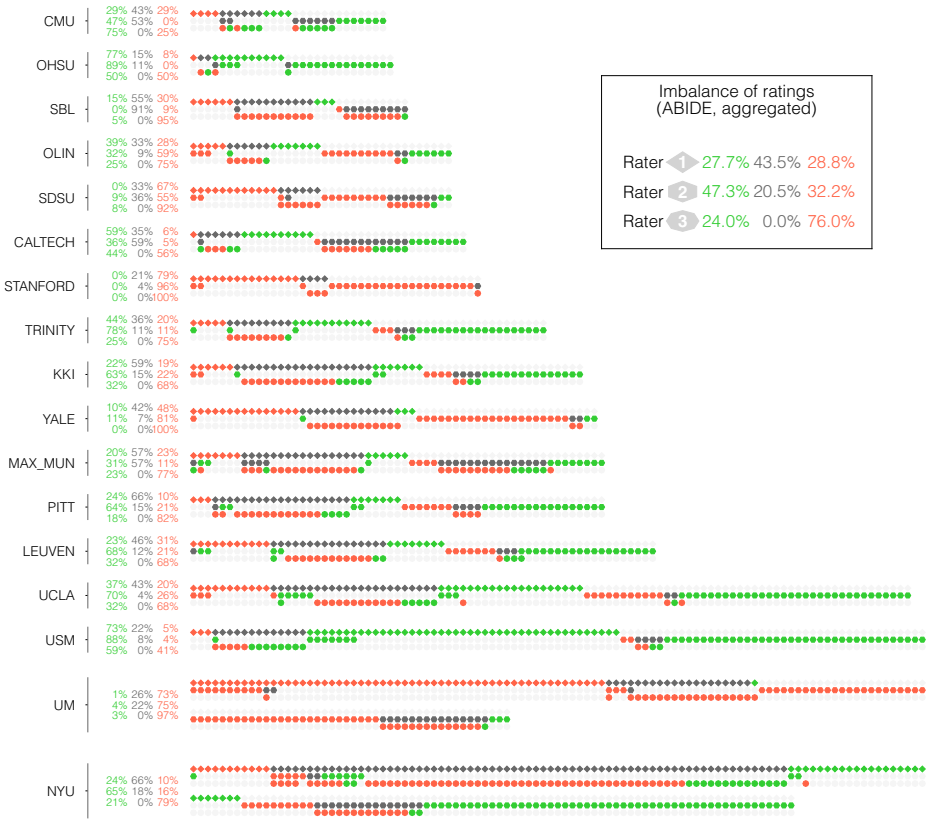

Fig. 2. Manual quality assessment of the ABIDE dataset by three raters. This figure shows the high inter-rater variability when assigning quality annotations ("accept"-green, "doubtful"-gray, "exclude"-red), as showed by the inconsistencies between the overlapping data points in the first and second row of each site. The third row (represented by octagons) of each block contains the examples relabeled by rater 3 (OE) proposed in this paper.

3 Results

3.1 IQMs are highly structured around their site of origin

As presented in Figure 3, first column, the t-SNE projection of the features in the bidimensional space allows for easy visual clustering of data points by their site of origin. We also investigated the projection after three different manipulations of the input dataset. First, after site-wise normalization of samples (subtracting the mean of the corresponding site and dividing by the standard deviation of the site). Second, projection of the samples after the first feature selection step of our classifier, which removes 14 features highly predictive of the site. Finally, we plotted the t-SNE projection of the site-wise normalized features that survived the aforementioned selection. Only in this last condition, the structure of

the dataset with respect to the site of acquisition seems weaker, although still present.

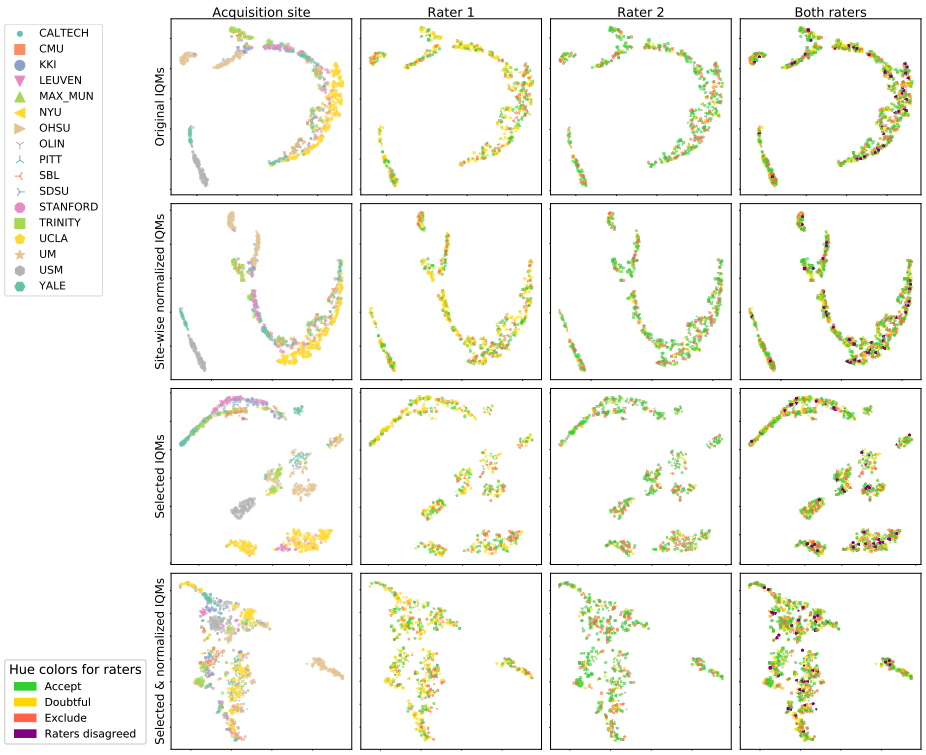

Fig. 3. T-SNE (t-distributed Stochastic Neighbor Embedding) of the examples belonging to the ABIDE dataset. While the dataset shows a strong structure with respect to the site of acquisition (first column), the t-SNE projection did not reveal any obvious bias within and across human ratings (three rightmost columns). Rows represent data manipulations: 1) features are used as calculated by MRIQC ("original IQMs"); 2) features are site-wise normalized to be zero-centered and have standard deviation of 1.0; 3) features that are highly predictive of the site of origin are removed; 4) the features selected in 3, but site-wise normalized.

3.2 T-SNE did not show obvious structure with respect to quality annotations

We could not identify any clear structure as regards to the two raters and/or the direction of their ratings, neither at the full-sample level (Figure 3) nor within-site (Figure 4). When inspecting the t-SNE projection of different scanning sites (Figure 4), ratings from both experts and their labels do not seem to cluster together.

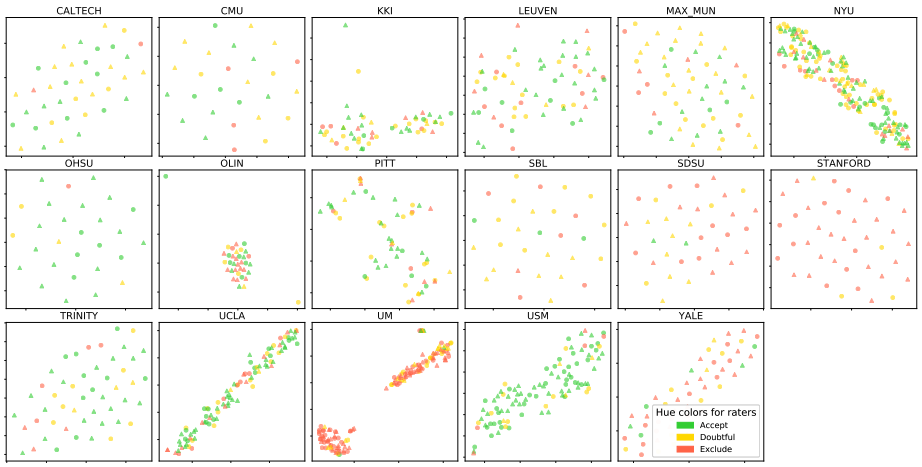

Fig. 4. The t-SNE projection per site did not show any particular clustering by rater (shape) nor their annotations (color coded).

3.3 Out-of-sample accuracy and sensitivity improve after denoising the quality annotations

A total of 362 examples from ABIDE and 155 from ds000030 were relabeled. After the process, the classification outcomes were not as imbalanced as the binarized versions of the classifiers anymore. The QC annotations for the ABIDE dataset changed from 63%/37% to 44.87%/55.13% ("accept"/"exclude"), and 83%/17% to 64.53%/35.47% for the case of ds000030. After revision, ∼24% of the ABIDE's "doubtful" instances and ∼82% of those in ds000030 were mapped to the "accept" rating. We then re-ran model selection and training on the ABIDE dataset, as described in our previous work [5]. Testing on the held-out dataset (ds000030) the performance (accuracy/area under the curve, ACC/AUC) increased from ∼72%/0.82 to ∼81%/0.85 (Table 1). The sensitivity to "exclude" cases increased from 0.20 to 0.67. Therefore, revising the labels improved the performance, and particularly the sensitivity to identify degenerate images.

Table 1. Effect of denoising on the performance attained by the model.

		precision	recall	f1-score	accept	exclude	total
Before denoising	accept	0.70	1.00	0.82	171	0	171
	exclude	1.00	0.20	0.34	75	19	94
	avg / total	0.80	0.72	0.65	246	19	265
		precision	recall	f1-score	accept	exclude	total
After denoising	accept	0.83	0.88	0.86	151	20	171
	exclude	0.76	0.67	0.71	31	63	94
	avg / total	0.80	0.81	0.80	182	83	265

4 Discussion and conclusion

In our original MRIQC proposal [5] we investigated the feasibility of automated quality prediction on T1w images based on a number of features (image quality metrics, IQMs). We discovered a strong relationship between the IQMs and the site where the images were acquired, which hampers generalization of trained models. We also analyzed the inter-rater variability of the quality annotations done by two experts and found that these labelings agreed moderately. Using t-SNE (t-student Stochastic Neighbor Embedding), we projected the IQMs extracted with MRIQC on a bidimensional space easier to visually interpret. Through t-SNE projections, we confirmed again the hypothesis that IQMs are strongly structured by their site of origin (see Figure 3). The strong relationship between IQMs and the acquisition site and parameters poses serious difficulty on generalization of learning approaches. We also investigated the quality annotations by experts using the t-SNE analysis, and could not confirm similar structures as regards raters and/or their ratings (see Figure 3). The same t-SNE analysis conducted site-wise did not reveal any clear relationship between IQMs and features, suggesting that, i) better features more representative of quality should be defined; and ii) the problem is not reliably solved with simple models.

In our prior work, two independent experts rated 601 and 866 images respectively, corresponding to 1367 T1-weighted images that belong to the ABIDE dataset and OpenNeuro's ds000030 dataset. A subset of 100 images were rated by both experts to test intra-rater reliability. They assigned three quality categories (namely "accept", "doubtful", and "exclude") to each example. We binarized the quality ratings by mapping all "doubtful" examples to "accept" instances. We revise such decision here, and, in order to rule out the binarization as a factor penalizing performance, we conducted a relabeling of all "doubtful" examples by a third expert (OE). Then, we repeated the model selection and evaluation process described in [5], and report an improvement on accuracy from 72% to 81% and an improvement on sensitivity to the "exclude" class from 0.20 to 0.67. We interpret that the binarization step mapping "doubtful" examples to the "accept" class introduced labeling errors that were larger for the ABIDE dataset. Of all "doubtful" instances in our test and train sets, ∼24% of the train set (ABIDE) and ∼82% if the test set (ds000030) were mapped to the "accept" rating. That is in contrast to the naive decision we had originally made mapping all instances to the "accept" rating. Particularly, for ABIDE, the "majority" mapping that would have introduced less errors would have assigned the "exclude" class for all "doubtful" examples. Such imbalance, and the opposite directions of the "majority" mapping reflects the fact that ds000030 was rated by only one of the original raters, and they conducted this rating before any assessment of ABIDE. This reveals the need for expert "calibration" since "doubtful" examples in ABIDE generally showed lower quality than those in ds000030 to the eyes of the third rater. Also, the labeling protocol also arises as a great source of variability as we already had tested in [5]. In this case, the revision of ratings was conducted on MRIQC-generated visual reports, which is different from the protocol used for the original labeling.

A limitation on the interpretation of the performance improvement sources from the fact that the rater who revised the quality annotations had already seen many predictions on ds000030 done with the former classifier. Although we did not train the classifier on ds000030, this knowledge could have been leaked into the training set through the relabeling of ABIDE. Therefore, a new, independently rated dataset would be necessary to rigorously confirm that the improvement on performance we found is not a consequence of backpropagating knowledge through the rater.

This work demonstrated the challenging problem of "site-effects" in the IQMs extracted with MRIQC from images of our test set. We also demonstrate how reducing errors in the labels of both training and held-out datasets accounted for a ∼10% improvement on accuracy, and a ∼0.46 increment in the sensitivity to the "exclude" class. We conclude that MRIQC's classifier could be further improved along several directions: i) the collection of a larger training set with labels annotated by several experts to allow denoising them with averaging or majority voting; ii) the conversion of the classification problem into a regression problem so the prediction error is not penalized by binarization; and iii) the addition of more IQMs that are less prone to "site-effects".

Bibliography

[1] Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* **59**, 2142–2154 (2012). doi: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* **88**, 79–90 (2014). doi:10.1016/j.neuroimage.2013.11.027.

[3] Reuter, M. *et al.* Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage* **107**, 107–115 (2015). doi:10.1016/j.neuroimage.2014.12.006.

[4] Alexander-Bloch, A. *et al.* Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human Brain Mapping* **37**, 2385–2397 (2016). doi:10.1002/hbm.23180.

[5] Esteban, O. *et al.* MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE* **12**, e0184661 (2017). doi:10.1371/journal.pone.0184661.

[6] Di Martino, A. *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry* **19**, 659–667 (2014). doi:10.1038/mp.2013.78.

[7] Poldrack, R. A. *et al.* A phenome-wide examination of neural and cognitive function. *Scientific Data* **3**, 160110 (2016). doi:10.1038/sdata.2016.110.

[8] Gorgolewski, K. J., Durnez, J. & Poldrack, R. A. Preprocessed Consortium for Neuropsychiatric Phenomics dataset. *F1000Research* **6**, 1262 (2017). doi:10.12688/f1000research.11964.2.

[9] Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**, 733–739 (2010). doi:10.1038/nrg2825.

[10] Woodard, J. P. & Carley-Spencer, M. P. No-Reference image quality metrics for structural MRI. *Neuroinformatics* **4**, 243–262 (2006). doi:10.1385/NI:4:3:243.

[11] Mortamet, B. *et al.* Automatic quality assessment in structural brain magnetic resonance imaging. *Magnetic Resonance in Medicine* **62**, 365–372 (2009). doi:10.1002/mrm.21992.

[12] Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* (2017). doi:10.1016/j.neuroimage.2017.10.034.

[13] Pizarro, R. A. *et al.* Automated Quality Assessment of Structural Magnetic Resonance Brain Images Based on a Supervised Machine Learning Algorithm. *Frontiers in Neuroinformatics* **10** (2016). doi:10.3389/fninf.2016.00052.

[14] Shehzad, Z. *et al.* The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of MRI data. In *Front. Neurosci. Conference Abstract: Neuroinformatics 2015* (Cairns, Australia, 2015). doi:10.3389/conf.fnins.2015.91.00047.

[15] Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).