Defacing biases visual quality assessments of structural MRI

Céline Provins[1], Élodie Savary[1], Thomas Sanchez[2,1], Emeline Mullier[1,3], Jaime Barranco[2,1,4,5], Elda Fischi-Gómez[2,1,6], Yasser Alemán-Gómez[1,7], Jonas Richiardi[1,2], Russell A. Poldrack[8], Patric Hagmann[1], Oscar Esteban[1]

[1]Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[2]CIBM - Center for Biomedical Imaging, Switzerland

[3]EEG and Epilepsy Unit, University Hospital and Faculty of Medicine of Geneva, University of Geneva, Switzerland

[4]School of Engineering, Institute of Systems Engineering, HES-SO Valais-Wallis, Sion, Switzerland

[5]The Sense Innovation and Research Center, Lausanne and Sion, Switzerland

[6]Signal Processing Laboratory (LTS5), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[7]Center for Psychiatric Neuroscience, Department of Psychiatry, Lausanne University Hospital and University of Lausanne. Prilly, Switzerland.

[8]Department of Psychology, Stanford University, Stanford, CA, US

Abstract

A critical requirement before data-sharing of human neuroimaging is removing facial features to protect individuals' privacy. However, not only does this process redact identifiable information about individuals, but it also removes non-identifiable information. This introduces undesired variability into downstream analysis and interpretation. This registered report investigated the degree to which the so-called *defacing* altered the quality assessment of T_1-weighted images of the human brain from the openly available "IXI dataset". The effect of defacing on manual quality assessment was investigated on a single-site subset of the dataset (N=185). By comparing two linear mixed-effects models, we determined that four trained human raters' perception of quality was significantly influenced by defacing by comparing their ratings on the same set of images in two conditions: "nondefaced" (that is, preserving facial features) and "defaced". In addition, we investigated these biases on automated quality assessments by applying repeated-measures multivariate ANOVA (rm-MANOVA) on the image quality metrics extracted with *MRIQC* on the full IXI dataset (N=581; three acquisition sites). This study found that defacing altered the quality assessments by humans and showed that *MRIQC*'s quality metrics were mostly insensitive to defacing.

Introduction

The removal of facial features —or *defacing*— is necessary before sharing anatomical images of the brain to protect participants' privacy[1] in compliance with some local privacy protection regulations, such as the General Data Privacy Regulation[2] (GDPR) in Europe or the Health Insurance Portability and Accountability Act[3] (HIPAA) in the US. Defacing is typically implemented by zeroing, shuffling, or filtering the content of image voxels located in an area around the participant's face and, often, the ears (see Fig. 1). Defacing is a destructive step with the potential to alter the results of downstream processing. De Sitter et al.[4] showed that downstream automated analysis methods failed in execution up to 19% of the cases after defacing, as opposed to 2% on nondefaced counterparts. They also reported systematic differences between the same processing with and without defacing in several outcomes of interest in neurodegeneration studies. Schwarz et al.[1] showed how these failures propagate and accumulate downstream, leading to substantial changes in the study outcomes. In a similar approach to our design, Bhalerao et al.[5] explored the impact of different defacing tools on a subset of image quality metrics (IQMs) automatically generated with *MRIQC.*[6] They found that all defacing tools impacted a subset of IQMs, and they estimated corresponding effect sizes on a sample limited to 30 subjects with a univariate modeling approach. Moreover, they identified further effects on the downstream segmentation of images. To the best of our knowledge, no study has investigated the impact of defacing on the manual assessment of MRI data quality.

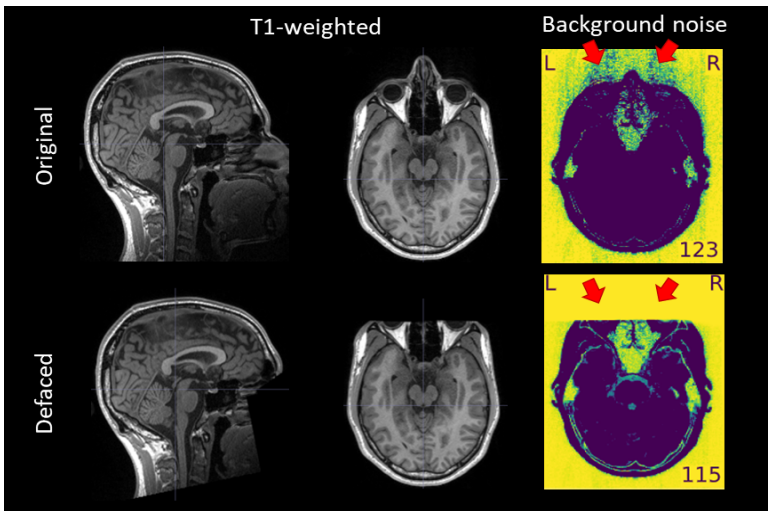

Figure 1. An example of T1w image before and after defacing. Defacing is typically implemented by zeroing the voxels around the face. The background noise visualization is extracted from the *MRIQC* visual report and illustrates that eye spillover is one example of information key in evaluating image quality that is removed by defacing.

Understanding how defacing alters the initial quality assessment (QA) checkpoint of the research workflow[7] is necessary because there is compelling evidence that data showing specific artifacts or insufficient overall quality introduce bias into the results of analyses, raising questions about their validity.[8–10] For example, Alexander-Bloch et al.[10] showed that in-scanner motion can lead to systematic and region-specific biases in anatomical estimation of features of interest, such as cortical thickness. Therefore, it is critical to reliably identify substandard MRI data and exclude them early from the research workflow (quality control, QC). However, the reliability of automated QC alternatives is limited mainly due to site-effects[11], which leads to implementing QA manually by screening the imaging data on a one-by-one basis. Because visual inspection is time-consuming and prone to large intra- and inter-rater variabilities, implementing assisting tools and protocols to screen and QA large datasets efficiently is an active line of work (e.g., *MRIQC*[6], *MindControl*[12], and *Swipes4science*[13]). Large consortia have also made substantial investments in this critical task and have generated valuable contributions to QA/QC protocols, e.g., the Human Connectome Project[14] or the INDI initiative (*QAP*[15]). One related and conceptually innovative approach was proposed by the UK Biobank[16], where sufficient quality was operationalized as the success of downstream processing. Given the massive size of the UK Biobank, Alfaro-Almagro et al.[16] flagged the images that did not successfully undergo pre-processing for exclusion. Although image exclusions were related most often to qualitative issues on images (e.g., artifacts), some images were discarded without straightforward mapping to quality issues. Moreover, because the QA/QC is onerous, many teams have attempted automation, either by defining *no-reference* (that is, *no ground truth is available*) IQMs that can be used to learn a machine predictor[6,17,18], or by training deep models directly on 3D images[19]. However, the problem remains extremely challenging when predicting the quality of images acquired at new MRI devices yet unseen by the model[6,11].

In a recent exploration[20], we found preliminary evidence that defacing alters both the manual and automatic assessments of T_1-weighted (T1w) MRI on a small sample (N=10 subjects) implemented with *MRIQC*. The present registered report builds on our exploratory analysis to confirm that four

trained human raters make systematically varying QA decisions when rating the same images depending on whether they have been defaced. Regarding automated QA, this registered report shows that differences in IQMs extracted with *MRIQC* before and after defacing are not meaningful.

Methods

Hypotheses

This pre-registered report was set out to confirm whether defacing altered the manual and automatic QA of T1w images of the healthy human brain —implemented with *MRIQC*. This overarching question was tested in two specific hypotheses:

1. Defacing influences trained experts' perception of quality, leading to significant differences in quality ratings between the defaced and the nondefaced images. Specifically, we expected raters to assign higher ratings —on average— in the defaced condition than in the corresponding nondefaced condition (see Fig. 1).

2. Defacing influences automatic QA/QC with *MRIQC*, introducing significant and systematic biases in vectors of IQMs extracted from defaced and nondefaced images. As evidenced by our preliminary data[20], these biases may showcase one direction for some IQMs and the opposite or no effects on others. Therefore, a directionality of effects could not be hypothesized.

Data

This confirmatory analysis was based on the publicly available IXI dataset[21], which contains 581 nondefaced T1w images acquired at three different sites in London (UK). One site was the Hammersmith Hospital (HH, in the following), which used one 3 Tesla (3T) scanner. The two other sites (Guy's Hospital and Institute of Psychiatry, Psychology & Neuroscience; GH and IoPPN, respectively) featured 1.5T devices. The scanner parameters available for each site are listed in Table S1. None of the authors had screened or queried the dataset before pre-registration to anticipate quality-related patterns or summary statistics. Moreover, except for author OE, authors neither accessed nor performed any processing on the data before pre-registration. To investigate Hypothesis 1, the initial inclusion criterion was "all subjects with a T1w image acquired at the 3T site (HH)", yielding a total of 185 subjects. In testing Hypothesis 2, we included all participants in the IXI dataset with a T1w scan (581 subjects). The local ethics committee approved the processing of nondefaced images. This study did not attempt to re-identify the participants nor facilitate such efforts in any way.

Data processing. First, a defaced version of each scan (N=581) was generated with *PyDeface*[22]. We chose *PyDeface* because it presents the highest success rate at removing facial features while not removing brain voxels[23]. Furthermore, Bhalerao et al.[5] showed that *PyDeface* resulted in the smallest effect size on a subset of *MRIQC*'s IQMs. Fatal failure of *PyDeface* was the only —implicit, as we did not explicitly state fatal failure as an exclusion criterion due to oversight— exclusion criteria. At pre-

registration, we did explicitly specify that if an image were not fully defaced —meaning it partially retained facial features such as sections of the eyes and the surrounding background)— we would not exclude it. Under our hypotheses, such images were expected to be more consistent between conditions for humans and machines. Nonetheless, we decided to retain them despite their potential contribution to reducing effect sizes. No image led to the fatal failure of *PyDeface* nor retained partial facial features. The latest version in the 23.1 series of *MRIQC* (version 23.1.0) was then executed on all the T1w images available (that is, nondefaced and defaced). Then, the IQMs generated by *MRIQC*, corresponding to every image in the two nondefaced/defaced conditions, were collated and converted into a tabular format. Two subjects from the GH site (1.5T) were excluded because *MRIQC* failed to run on both the nondefaced and defaced images, reducing the dataset size to test Hypothesis 2 to 579 subjects. Hypothesis 1 (HH site) was not affected by this exclusion.

Manual assessment protocol. We performed manual quality assessment only on the images from the site with a 3T device (HH; 185 subjects). This choice eliminated the field strength and other variability sources emerging from the specific scanning site as potential random effects. Moreover, images acquired with the 3T scanner were expected to showcase a signal-to-noise ratio (SNR) approximately twice as high as the SNR of images acquired with 1.5T scanners. Thus, the images acquired with the 3T scanner likely yielded, on average, better quality assessments by human raters independently of the defacing condition. Four human raters (authors TS, EM, JB, and EFG, incorporated into the report after data collection) assessed the quality of the subsample in each of the two conditions (that is, defaced and nondefaced). The quality assessment was carried out with the individual screening of one *MRIQC*-generated visual report per subject and condition. Raters did not have access to the mapping between defaced and nondefaced counterparts. We concealed participant identifiers and shuffled their ordering before presentation. These reports are openly shared (see Data and code availability statement). Raters were recruited by inviting researchers who demonstrated an interest in QA within the Department of Radiology of the Lausanne University Hospital (CHUV, Lausanne, Switzerland) via e-mail. We did not impose restrictions on the raters' experience beyond familiarity with T1w images of the human brain. To ensure consistency of their training, raters read our published QC protocol[24] and took a 3-hour training session. The training session included an introduction to *MRIQC*'s visual reports, its "rating widget" (Fig. S1) as well as *Q'kay*[25]. *Q'kay* is a Web Service we developed for this study to assign and record the quality ratings. We also presented the list of quality criteria to consider, and for each criterion, we illustrated an exclusion example. Lastly, in the same session, the raters were asked to rate 20 images independently, and we compared and discussed the ratings together to ensure that the quality criteria were interpreted consistently. The images presented in the training session were extracted from diverse publicly available datasets[26–29] as well as in-house datasets. Those datasets were selected to illustrate all quality criteria and present a wide range of quality. None of the training examples were extracted from the IXI dataset. At the end of this session, the raters self-assessed their experience as either beginner, intermediate, or advanced. The materials corresponding to the

training session as well as the self-assessments of experience, are openly shared for future exploration (see Data and code availability statement). To assess the intra-rater effects on QA, 40 subjects selected randomly were presented a second time in both conditions to all raters without their knowledge. This summed up to 450 images per rater (225 images per condition). We chose to repeat 40 subjects because it represented a good trade-off between having enough statistical power and the risk of having raters who would not complete their assignment. Both the random number generator to choose the 40 repeated subjects and the blinding of participant identifiers were initialized with the timestamp of the registered report submission toward Stage 1. It was converted to an integer with the format YYMMDD + SSmmHH (Y: year, two last digits; M: month, D: day; S: seconds; m: minutes; H: hour). This seed was preserved, clearly reported, and set for all the analyses. After screening each visual report, the raters assigned each image a quality grade with the rating widget in Fig. S1. A quality score was assigned using a slider that permitted the selection of numbers on a continuous scale from 1.0 to 4.0 (interval step of 0.05 and the value of 1.0 corresponding to the lowest quality). As presented in Fig. S1, the slider was presented with four categorical ranges for reference. The starting position of the slider was set in the middle. The raters were instructed to assess each subject according to our QC protocol[24], and they did not have access to the IQMs. All raters viewed the visual reports on a single LED panel of 43" screen diagonal, 3840×2160 resolution, a typical static contrast of 5,000:1, and the same ambient lighting. *MRIQC* reports feature a stopwatch that records the exact time each assessment took. The time for each assessment was measured and is available for future exploration (see Data and code availability statement). The removal of the IQMs from the visual report and the tracking of raters' progress were managed with *Q'kay. Q'kay* also handled the order of presentation and assessment of the reports, ensuring it was the same for all raters. Our pre-registered methods stated that images would be excluded from all experiments when assigned the lowest grade (one in our 1-4 interval scale) in both conditions by all raters. No images were excluded from our analyses based on this particular exclusion criterion.

Experiments

Determining that defacing biases the human raters' assessments of quality. We tested the influence of the defacing condition and the rater (within-subject factor variables) on the ratings (dependent variable) using repeated-measures ANOVA (rm-ANOVA). In case data did not meet the rm-ANOVA model's assumptions regarding normality and sphericity, we had pre-registered a comparison between linear mixed-effects (LME) models. As opposed to multiple *t*-tests, rm-ANOVA and LMEs enable disentangling the inter-rater variability from intra-rater variability due to defacing and quantifying the latter. Because we did not necessarily expect the ratings' distribution of each rater to have the same mean, rm-ANOVA and LMEs accounted for the baseline difference in ratings by adding the rater as a random effect in the model. Normality was verified with the Shapiro-Wilk normality test[30], employing the *shapiro.test* function of the *ggpubr R* package[31,32]. Sphericity was assessed with Mauchly's test for sphericity[33] within the *rstatix R* package[32,34]. rm-ANOVA was

implemented with the *anova_test* function (*rstatix*; p<.02). We had preregistered a sensitivity analysis (Table 1) showing that our experimental design with rm-ANOVA could identify effects of size f=0.218 / η_p^2=0.045 (Table 1) with a power of 90%. For context, we found an effect size of f=0.31 (see Equations S3 and S4) in our pilot study. Comparison between both effect sizes needs, however, to be performed cautiously as the design of the rating collection has been modified between the pilot study and this pre-registration. In the contingency that at least one of the assumptions of rm-ANOVA was violated, we planned to compare two LMEs instead, one "alternative" model with defacing as a fixed effect and one "baseline" without defacing as a factor. In both models, the intercept was allowed to vary between raters (i.e., that factor was added as a random effect). The models were implemented in *R* with the *lmer* function of the *lme4* package[36]. As part of regression diagnostics (e.g., non-Gaussian or heteroscedastic residuals indicate non-optimal model fit), we examined the shape of the regression residuals, which are reported in the supplementary materials for completeness. To test the effect of defacing, we performed a likelihood-ratio test comparing baseline and alternative LMEs. The likelihood-ratio test was implemented with the *anova* function of the *R* package *stats*[32]. We set the significance threshold for the likelihood ratio at *p*=0.02. We pre-registered a sensitivity analysis by computing the noncentrality parameter λ associated with the likelihood ratio test, which is a proxy for its power[37]. This analysis indicated that $\lambda \geq 13.017$ would allow the detection of differences between models at 90% power (see Table 1). Lastly, the variance related to the intra-rater effect was estimated using the rm-ANOVA or, in the contingency case, by computing the variance of the regression coefficients linked to the random effect.

Table 1. Sensitivity analyses for rm-ANOVA and LME comparisons. We determined that our rm-ANOVA modeling would confirm differences in manual ratings of f=0.218 or larger with *G*Power*[35]. This sensitivity corresponds to η_p^2 =0.045 (i.e., a medium effect size) following Equation S2. In the rm-ANOVA sensitivity analysis, we set two groups (defaced/nondefaced) and four measurements (4 raters) with a total sample size of 185 subjects from the HH site, 90% power, α = 0.02, a nonsphericity correction of 0.34, and a correlation among repeated measures of 0.1. Note that this sensitivity analysis is conservative as we expected a much higher correlation among repeated measures, which would reduce the detectable effect size. Remaining conservative, we iteratively tried different non-sphericity correction values and kept the lowest one possible to maximize the detectable effect size. With *G*Power* (Fig. S7), we also estimated the noncentrality parameter λ associated with the likelihood ratio test, which is a proxy for the effect size, yielding λ = 13.017. The degrees of freedom of the likelihood ratio test correspond to the difference in parameter count between the two nested LMEs compared (see Table S6).

Test	Input parameters		Output parameters	
rm-ANOVA	α err prob	0.02	Noncentrality parameter λ	13.273
	Power (1-β err prob)	0.9	Critical F	5.454
	Total sample size	185	Df (Numerator)	1.02
	Number of groups	2	Df (Denominator)	186.66
	Number of measurements	4	Effect size f	0.218
	Corr among rep measures	0.1		
	Non-sphericity correction ε	0.34		
Likelihood ratio test	α err prob	0.02	Critical χ^2	5.412
	Power (1-β err prob)	0.9	Noncentrality parameter λ	13.017

	Degrees of freedom	1

Inspecting whether ratings are higher on defaced images. We used Bland-Altman (BA) plots[38] to visualize the bias and the limits of agreement of manual quality ratings between the nondefaced and the defaced conditions. Five BA plots were generated (Fig. 2), one for each individual rater and one pooling the ratings from all raters together. We used the individual-rater BA plots to investigate whether the bias varied with respect to the quality grade attributed and how the bias changed depending on the rater. The BA plot with the pooled ratings was used to test the significance of the bias. To demonstrate that the ratings of the defaced condition are higher than the corresponding ratings on the nondefaced condition, the bias should be shown to be significantly negative. Biases in the BA plot were deemed significant if the zero difference fell outside the 95% limits of agreement. We used non-parametric 95% limits of agreement because the paired differences were not normally distributed[39].

Determining whether defacing biases *MRIQC*-generated IQMs. Defacing impact on automatic QA was evaluated based on 58 out of 62 IQMs calculated by *MRIQC*. Two IQMs, Mortamet[18] "quality index 1" and the "5th percentile of the background intensity", were excluded because they were zero for all images. Two additional IQMs, the median and the median absolute deviation of the background intensity, were excluded as they were constantly zero except for two images. The distributions of all IQMs are plotted in the supplementary Fig. S13, with the plots corresponding to the four excluded IQMs highlighted in red. For the complete list of IQMs produced by *MRIQC*, please refer to Esteban et al.[6]. A two-way rm-MANOVA was used to test whether defacing significantly influenced the IQMs. This test was implemented with the *multRM* function of the *MANOVA.RM* package in *R*[32,40]. To account for the fact that the IQMs were heavily correlated (see Fig. S15), their dimensionality was reduced using principal component analysis (PCA) before applying rm-MANOVA. To ensure the defacing effects were not mitigated, PCA was fitted only on the IQMs extracted from the nondefaced data, and the resulting transformation was applied to the entire dataset, regardless of their nondefaced or defaced provenance. PCA was implemented with the *prcomp* function of the stats package of *R*, with the option *scale=TRUE*, meaning that the variables were standardized to have unit variance before the decomposition. The number of principal components was determined by the Kaiser criterion, which keeps components with an eigenvalue above 1.0. The rm-MANOVA was constructed with the projected IQMs as the continuous dependent variables and two categorical independent variables, one corresponding to the (nondefaced or defaced) condition of the image and the other corresponding to the scanning site. Adding the scanning site as an independent variable allowed us to control for site effects[11,41]. We applied a significance level of $p<.02$ for the rm-MANOVA and considered the p-values extracted under the Wald-type statistics section. We determined that our experimental design could identify, with a 90% power, effects of f = 0.16 / η_p^2 = 0.025 (i.e., a medium effect) or greater with *G*Power*[35] (see Table 2). For context, the effect size associated with the MANOVA on the IQMs of our pilot study was f=0.16 (see Equation S5 and S6 for its

computation). Comparison between both effect sizes needs, however, to be exercised with caution as the statistical designs are different; in our pilot study, we used a standard MANOVA on only 5 IQMs that showed the strongest bias on the BA plot. In contrast, in this pre-registration, we planned to use a repeated-measures MANOVA with all IQMs projected onto the PCA basis. In addition, the effect size associated with Bhalerao et al.[5] ranged from f=0.045 to f=1.79 with a mean effect size across IQMs of f=0.61 (see Equation S4 for the conversion of Cohen's d to Cohen's f). To visualize the defacing bias on the automatic quality ratings, we also generated a BA plot (as described above) for each IQM and each principal component. All BA plots are reported in Figs. S16 through S19 in the Supplementary Materials.

Table 2. The sensitivity analysis indicated that the rm-MANOVA was able to confirm differences in IQM of f=0.16 corresponding to $\eta^2 = 0.025$ (i.e a medium effect) or greater. We ran the sensitivity analysis with *G*Power*, setting three groups (3 sites) and two measurements (defaced/nondefaced) with N = 580 (number of T1w per subject) per condition, with 90% power, and α = 0.02.

Input parameters		Output parameters	
α err prob	0.02	Non centrality parameter λ	15.5188571
Power (1-β err prob)	0.9	Critical F	3.9386666
Total sample size	580	Numerator degree of freedom	2
Number of groups	3	Denominator degree of freedom	577
Number of measurements	2	Effect size f	0.1635746
		Pillai V	0.0260594

Methodological precisions and adaptations at Stage 2

An alternative to the standard BA plot is presented (Hypothesis 1). The BA plot is designed for test-retest reliability assessment, in which neither of the two measurements being presented can be assumed to be more accurate. However, in our design, ratings on the nondefaced condition can be assumed to be the reference value from which then test deviation in the defaced condition. Therefore, we modified the BA so that the x-axis represents nondefaced ratings (in contrast to the average between defaced and nondefaced of the "standard" plot). This modification of the BA plot displays the directionality of biases introduced by defacing more clearly. A side-by-side comparison between the standard BA plots and our "optimized" BA plot is shown in Fig. S2. Our variation of the BA plot revealed that images rated highly in the nondefaced condition were, by design, very unlikely to receive a higher rating in the defaced condition, which we did not anticipate at pre-registration time.

Visually assessing homogeneity of distributions across raters (Hypothesis 1). We visually explored the sample before carrying out the rm-ANOVA and LME tests. We evaluated the agreement across raters by generating "test-retest violin plots", where the left part of the violin shows the distribution of ratings in the nondefaced condition, and the right side shows the distribution of ratings in the defaced condition (Fig. 3).

Clarifications regarding the control for multiple comparisons (Hypothesis 1). In addition to the pre-registered test, we repeated the rm-ANOVA and LMEs analyses with two subsamples. We additionally ran both models (likelihood ratio test and rm-ANOVA) including only images that were rated "poor" or "exclude" (that is, ratings below 2.45) for two reasons. First, our BA plots showed that nondefaced images that were highly rated would not increase their rating after defacing. Second, we considered that QA/QC is more concerned with sensitivity (i.e., minimizing false negatives or, in other words, reliably detecting cases where the ratings on the defaced condition are substantially more optimistic about quality). In the second subsample, we considered only Raters 01, 02, and 03 because of the stark difference in Rater 04's distributions displayed by the test-retest violin plots (Fig. 3). With these additional four tests, and considering the two tests originally planned, we corrected the *p*-values for multiple comparisons (six tests) controlling for false discovery rate[42] (FDR).

The subject was added as a random effect in the LMEs (Hypothesis 1). Our pre-registered plan failed to identify subjects as a random effect.

Post-hoc sensitivity analysis implementation. Our pre-registered plan committed to estimating the effect size of the maximum likelihood ratio test. The likelihood ratio test compares the ratio of likelihoods to a χ^2-distribution with the degrees of freedom (df) equal to the difference in parameter counts between the nested alternative and baseline models[43,44]. As such, a proxy for power is given by the noncentrality parameter λ. For each LME, we obtained an unbiased estimate of the noncentrality parameter[43] using Equation S1 with df = 1**.**

Accounting for site differences in the IQMs analysis (Hypothesis 2). In addition to the pre-registered analysis plan, the IQMs' standardization and PCA fitting were implemented site-wise instead of considering the three sites together. We made that decision after observing that the pre-registered rm-MANOVA did not show significant IQM differences between sites (see Table 4) despite some site differences being visible in Fig. S14. Therefore, we computed the mean and standard deviation of the IQMs extracted from the nondefaced images of one site and standardized both the IQMs computed from nondefaced and defaced images from the same site by subtracting that mean and dividing by that standard deviation. Correspondingly, the following three PCA models were fit on the nondefaced sample separately by site. First, considering only IQMs computed from nondefaced images, we inspected how many components are needed in each site according to the Kaiser criterion. We chose the maximum number of components for all PCAs independently of the site. Selecting a common number of components across sites was necessary as the rm-MANOVA requires a single number of dependent variables. We then re-ran the PCA separately per site with the chosen number of components considering IQMs from both conditions. The subsequent steps of the analysis remained unchanged from the pre-registered plan.

Results

We confirmed defacing biased quality assessments by human raters. A likelihood ratio test demonstrated that including defacing as a fixed effect (the alternative model) explained significantly more variance than the baseline model (see Table 3; $p_{FDR} = 0.0183$). This result supports that defacing significantly influenced the quality assessments issued by human raters on the same images. Fig. S10 (Supplementary Materials) shows the distribution of residuals for each fitted LME model, as well as the Q-Q plot[45] for the residuals and the subject and the rater as random effects. The residuals for both the baseline and alternative models were approximately normally distributed (see Fig. S10), with heavier left tails. The residuals distribution plot also suggested signs of heteroscedasticity in the lower ratings, although most points were still evenly distributed around zero. Table S8 reports the variance of the random effects coefficients and residuals.

Table 3. The linear mixed-effect models (LME) with defacing as a fixed effect explained significantly more variance than a "baseline" counterpart without defacing. Model comparison with a likelihood ratio test yielded a $p_{FDR} = 0.0183$ after FDR correction. Complete reporting of the pre-registered comparison and the additional exploratory analyses are provided in Tables S6 and S7 (Supplementary Materials).

Sample size	χ^2	df	$P(>\chi^2)$	p_{FDR}	p_{FDR} <.02
1480	6.283	1	0.0122	0.0183	*

We carried out the LME model comparison because the majority of the rating subgroups defined by the interaction between the defacing condition and rater identifier were non-normally distributed according to the Shapiro-Wilk normality test[30] (see Table S3; Fig. S9). Nonetheless, we also carried out the planned rm-ANOVA for completeness, and the results of the rm-ANOVA are presented in the Supplementary Materials (Table S4). Even though non-normality undermines the sensitivity of rm-ANOVA, these tests also showed statistically significant biases between the defaced and nondefaced rating conditions.

Following our pre-registered plan, we also conducted a post-hoc power analysis of the alternative vs. baseline LMEs comparison to estimate an effect size. We estimated the noncentrality parameter corresponding to our maximum likelihood ratio test $\lambda = 5.28$ (see Table S9). Therefore, $\lambda < 13.017$ indicates that the effect size was smaller than the minimum our previous sensitivity analysis found detectable at 90% power (Table 1). Our pre-registered plan explicitly stated, "We will deem the effect irrelevant if the latter parameter [noncentrality parameter of the χ^2-distribution] is smaller than 13.017, corresponding to the minimum power achievable from the sensitivity analysis." However, in preparing this final report, we understood these post-hoc power analyses give no basis for "deeming an effect irrelevant", as effect size estimates can be extremely biased[46]. We report this post-hoc power analysis following the pre-registration, but remark that it should be interpreted with the utmost caution.

Biases were larger in images rated as having lower quality in the nondefaced condition. Our variation of the BA plot (Fig. 2) and our test-retest violin plots contrasting defaced and nondefaced ratings (Fig. 3) revealed that nondefaced images rated the lowest showed larger quality gaps with respect to their defaced counterpart than highly-rated nondefaced images. Low ratings presented an evolution line with a higher slope, sometimes jumping one unit (that is, a category change in the appreciation of quality, for example, from “poor” to “acceptable”), while the difference amongst the highest ratings remained within 0.5 unit. Our exploratory models considering only low ratings further supported this interpretation by showing a higher effect size linked to defacing than that of the pre-registered model counterparts with all ratings (Table S4). These plots also reflected that raters were not equally affected by defacing. Our most experienced rater (Rater 01) presented the highest bias, defined as nondefaced minus defaced $\Delta_{ndef\text{-}def}$ = -0.12 (Table S2), while a more junior rater (Rater 04) barely modified their quality assessment ($\Delta_{ndef\text{-}def}$ = -0.01). The differences between nondefaced and defaced ratings by Rater 04 could only be small because most of their ratings were remarkably high as compared to the other raters (see Fig. 3). Our subsample rm-ANOVA and LMEs tests without Rater 04 supported that Raters 01, 02, and 03 showed better agreement, as shown by the larger rater effect size compared to our pre-registered model counterparts with all ratings (Table S4). The bias (nondefaced – defaced, $\Delta_{ndef\text{-}def}$) was consistently negative across all raters, implying that defaced images were, on average, scored with a higher rating. Yet, this bias was not significantly consistent because the 95% limits of agreement on the BA plot pooling all raters contained zero, as shown in Fig. S2.

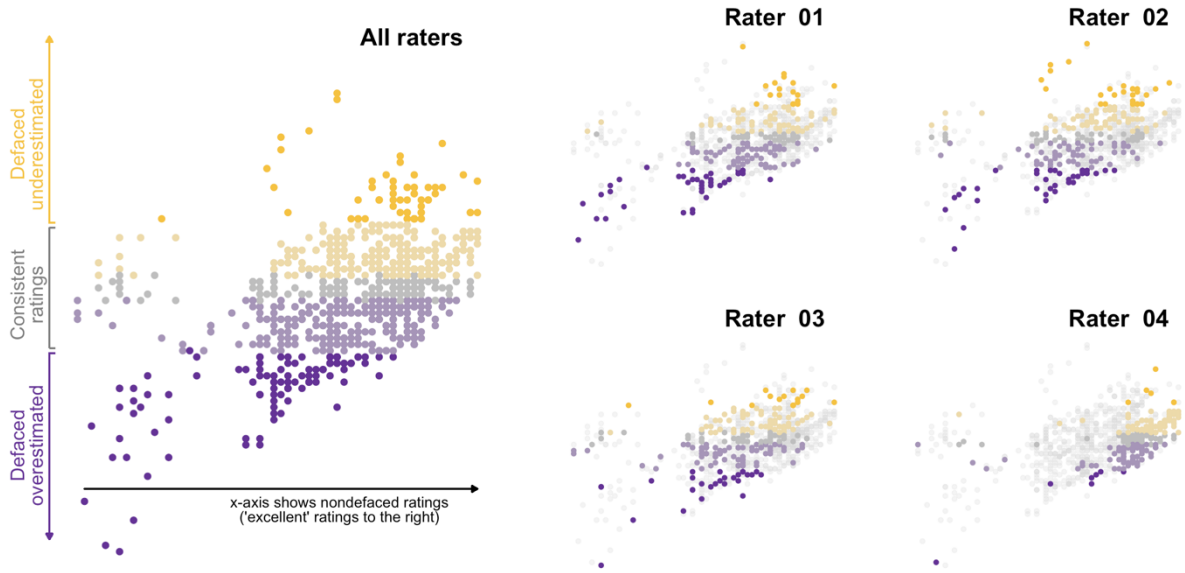

Figure 2. Defacing biases human assessment of image quality, particularly when image quality is low. Nondefaced images with low ratings were evaluated as similar or better quality after defacing, whereas nondefaced images with high ratings were assessed as similar or lower quality. This variability at the extremes is explained by our rating scale being constrained to the range [1,4]. More interestingly, we observe that lower ratings in the nondefaced condition vary the most. This suggests that when more pronounced artifacts were present, defacing removed critical information needed to evaluate image quality. The *y*-axis corresponds to the defaced rating subtracted from the nondefaced rating. Reliable ratings (i.e., ratings with absolute differences lower than 0.1) are colored in grey, and ratings that were overestimated or underestimated after defacing are colored in purple or yellow, respectively. The points are plotted with lower luminosity if their absolute difference is below 0.5, indicating a moderate defacing bias.

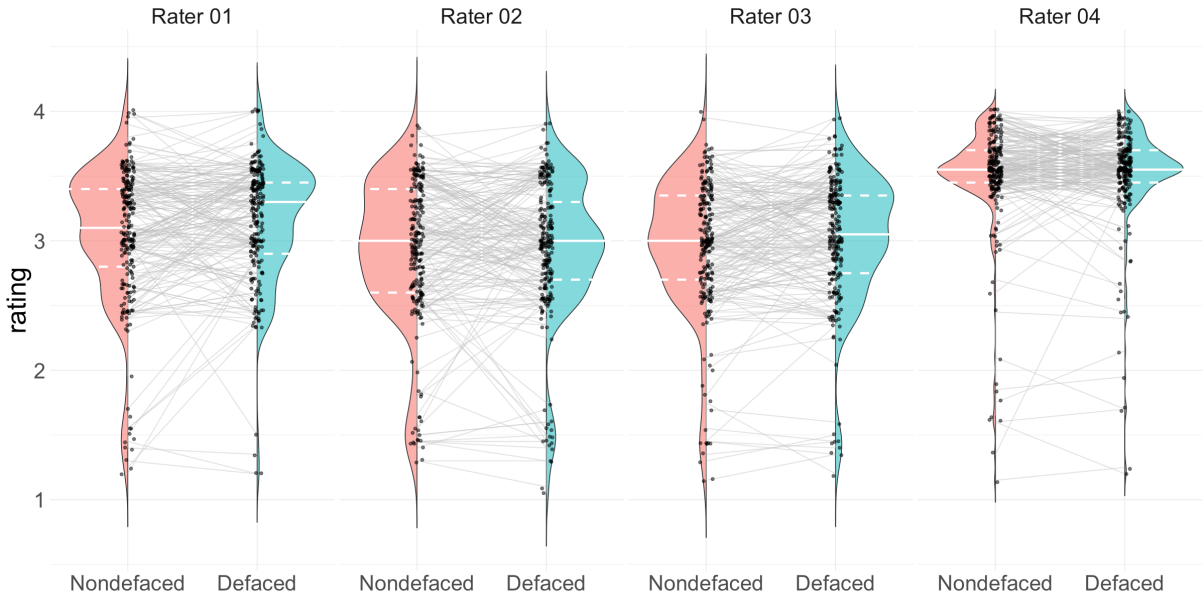

Figure 3. Rater 04 issued visibly different ratings, generally more optimistic, than the other raters in both (defaced and nondefaced) conditions. The grey lines highlight the evolution of the rating between the nondefaced image and its defaced counterpart. The full white line in the violin plot represents the median of the distribution, while the dashed white lines represent the 25% and 75% quantiles. Comparing the median of the rating distribution from the nondefaced versus defaced images, it is visible that different raters presented different bias magnitudes. Our most experienced rater (Rater 01) showed the largest bias. Rater 04's rating distribution diverged from that of the other raters, being more optimistic overall about the quality of the images. Rater 04 also displayed a lower spread in quality assessments (see Fig. S5). Lastly, low ratings tended to be more biased by defacing as they showed a steeper evolution line, sometimes jumping one unit or more (equivalent to switching categories in the appreciation of quality, e.g., going from "poor" to "acceptable"). Higher ratings displayed gaps within 0.5 units. The same observation is visible on the BA plots (Figs. 4 and S3).

***MRIQC*-generated IQMs were mostly insensitive to defacing.** Our pre-registered rm-MANOVA test (Table 4) did not reveal differences in the IQMs distribution before and after defacing (F(1, 184) = 0.003, p=0.96, η_p^2 = 3.5e-6) nor between sites (F(2, 184) = 0.376, p=0.828, η_p^2 = 7.9e-4). As explained in the section "Methodological precisions and adaptations at Stage 2", observing no site effects pushed us to refine our PCA approach. Following the Kaiser criterion, we kept ten components for each site (eigenvalues presented in Table S10). We also verified on a scree plot (see Fig. S20) that the components from the 11th did not explain a substantial part of the variance, confirming the choice of 10 components. The rm-MANOVA with refined PCA revealed a significant influence of defacing on the IQMs, although the effect size was very small (F(5,14)= 8.269, p< .001, η_p^2=0.0047, f=0.069; see Equation S6). In contrast, Bhalerao et al.[5] found bigger effect sizes for the impact of *PyDeface* on *MRIQC*-generated IQMs, with f ranging from 0.045 to 1.79 with a mean E[f] = 0.61. These post-hoc power analyses must be interpreted cautiously for the reasons stated above regarding our corresponding analyses within the LME comparison. In a post-hoc exploration using BA plots (Fig. S16), we found that only the entropy-focus criterion ("efc") presented a significant bias between the nondefaced and defaced images. This bias is also clearly visible in the violin plots (Fig. S14).

Table 4. Results of repeated-measures MANOVA on the projected IQMs. We did not find an effect of the site nor an effect of defacing on the principal components extracted from the IQMs. However, applying the IQMs standardization and PCA separately per site mitigated the site effects, revealing a defacing bias. The much smaller effect size associated with the defaced variable compared to the site variable in the rm-MANOVA, which included a single standardization and PCA across all sites, suggests that the larger confounding effect of the site likely overshadowed the impact of defacing. Despite being significant, the defacing bias is associated with a negligible effect size. Effect size is reported as partial eta-squared (η_p^2) and was computed with the function *F_to_eta2* from the *R* package *effectsize*[32,47], which implements the formula given in Equation S5 with df = numDF and df_error = denDF. p_{FDR} corresponds to *p*-values controlled for false discovery rate (FDR).

		Wald-type Statistic (WTS)				**Modified ANOVA-type Statistic (MATS)**		**Resampling version of the tests**	
		Test Statistic	df	p	p_{FDR}	Test Statistic	η_p^2	p_{WTS}	p_{MATS}
Single PCA (pre-registered)	site	0.376	2	0.828	0.994	0.457	$7.9 \cdot 10^{-4}$	0.829	0.827
	defaced	0.003	1	0.96	0.96	0.002	$3.5 \cdot 10^{-6}$	0.961	0.961
Site-wise PCA (exploratory)	site	0.358	2	0.836	0.994	0.710	0.0012	0.833	0.833
	defaced	181.182	1	<.001	<.003 **	2.726	0.0047	<.001 ***	<.001 ***

Discussion

We presented a registered report investigating the alterations introduced by defacing in the quality assessment of manual and automatic ratings of T1w whole-brain images. The automatic rating used image quality metrics (IQMs) —extracted with *MRIQC*— as a proxy. Our analysis for manual (human) ratings showed that a linear mixed-effects (LME) model with defacing as a fixed effect explained significantly more variability than a baseline model without defacing. However, the biases we estimated were relatively small: the largest bias $\Delta_{ndef\text{-}def}$ = -0.12 (Table S2) was displayed by the most experienced rater (Rater 01). This indicates that human raters are sensitive to defacing, possibly more for more experienced raters. Since defacing is necessary to protect participants' privacy, our work shows the importance of including QA/QC materials (such as *MRIQC*'s visual reports) extracted from the original (nondefaced) images when sharing defaced data, as long as the participant's identity cannot be reconstructed from those additional resources. This will enable data re-users to perform QA/QC in more adequate settings, minimizing the risk of artifacts going unnoticed and biasing analysis.

Visual examination of the manual ratings through optimized Bland-Altman (BA) plots displayed a phenomenon we should have anticipated before pre-registration, which we acknowledge as a limitation of the present work. In agreement with our hypothesis, nondefaced images rated with the lower quality ("exclude" through "poor" and "low-acceptable" regime, from 1 to 2.45 on our rating

scale) received substantially better ratings after defacing (see Fig. 2). However, when nondefaced images were assigned high ratings (i.e., “excellent”), it was very unlikely that the defaced image received a more optimistic rating because of the scale constraint. We failed to identify this trivial limitation, and therefore, future investigations should employ datasets with balanced representations of poor-quality and excluded images. This bias is present in all datasets (as QA attempts to minimize data exclusion on the grounds of insufficient quality), and it is more acute in public datasets (as data that met exclusion QC criteria are generally not shared openly as they were not considered for the analysis). For this reason, we argue that improving the automated QA/QC of T1w images will require that all available data be shared, including samples discarded because they met the study’s exclusion criteria. Doing so will enable the development of more sensitive and specific QA/QC tools that are robust to site effects. Sharing all data is also essential because an image deemed unsuitable for one study might be valuable for another, as exclusion criteria vary based on the specific project’s needs. Factors like the type of analysis or the region of interest in addressing the research question influence these QA/QC decisions[24].

While we designed the experiment with four raters, and despite the (pre-registered) efforts to calibrate the raters with a training session, one of the raters’ assessment distributions differed notably from the other three raters in location and spread. Their ratings concentrated within the “acceptable-to-excellent” range. We explored all models after discarding this rater, and alongside visualization through BA plots and “test-retest violin plots”, we found more reliable results (and larger effect sizes) excluding this rater. The distinctive distribution of ratings may indicate a failure in the design of the training session, leading to the rater’s application of starkly different assessment criteria. Our pre-registration failed to set corresponding QA/QC criteria to assess the raters’ performance and anticipate countermeasures. Therefore, it is critical to QA/QC protocols to assess raters’ calibration before assessment and then schedule checkpoints over time to take corrective measures and identify rater disagreement early. We share all the materials in this report to allow researchers to build upon these training and calibration materials. Nonetheless, it cannot be ruled out that other circumstances determined the differences in rating distribution, such as attrition drifts (Rater 04 evaluated most of the 450 images in a single session; Fig. S5), differences in line-of-sight angle, or individual behavioral variables (e.g., sleep deprivation).

We carried out several exploratory analyses in addition to the pre-registered confirmatory tests (reported as Supplementary Materials). For instance, we estimated, using the intra-class correlation coefficient, that our four raters presented an inter-rater reliability of 0.542 (95% CI = [0.497, 0.587]), corresponding to a moderate agreement (see subsection S4.2). With a subsample of 40 subjects randomly selected, we ran an exploratory analysis investigating the test-retest reliability of the raters in both nondefaced and defaced conditions. This analysis revealed good intra-rater agreements (Table S2), with varying and slight differences across raters regarding whether they became more or less critical of perceived quality over time (Fig. S5 and S6). Beyond identifying one rater’s

assessment sticking out compared to the other raters, our study was admittedly not designed to investigate how raters may be differently influenced by defacing (for example, in function of their experience), which remains a future line of research.

Finally, we also found that *MRIQC*'s IQMs are mostly insensitive to defacing. We tested the influence of defacing on the IQMs using rm-MANOVA after dimensionality reduction via PCA. Defacing bias became significant only when we meticulously controlled for site effects by applying standardization and PCA per site, and even then, it was associated with a minimal effect size η_p^2=0.0047. In a post-hoc exploration using BA plots, we identified the entropy-focus criterion ("efc") as the single metric that differed significantly after defacing. Our conclusion contrasts with Bhalerao et al.[5], who stated that *MRIQC*-generated IQMs are significantly influenced by defacing. One explanation for the disagreement is that they evaluated different defacing methods, and *PyDeface* —our choice— yielded the smallest influence on the IQMs. Secondly, they performed univariate tests on only ten IQMs, including "efc" and foreground-background energy ratio ("fber"), which were the IQMs varying the most with defacing when visually assessed (Fig. S14). These reasons explain why they found significant differences with a much smaller sample size. Based on our findings, we recommend that *MRIQC* (and any QA/QC tool for unprocessed T1w images) include metrics sensitive to artifacts in areas typically removed or altered by defacing (e.g., eye spillover)[48]. These metrics should be clearly identified and excluded from analyses involving defaced images where they are not available. Even though *MRIQC*'s IQMs seemed insensitive to defacing, we also carried out a reliability analysis of the IQMs to determine whether defacing altered the repeatability of IQM calculation (Supplementary Materials, subsection S6.4), concluding that defacing did not have an appreciable impact on IQMs reliability.

This registered report lays the groundwork for a better understanding of human expert behavior in the assessment of anatomical T1w MRI of the human brain. It provides a tooling framework to establish a rigorous QA/QC checkpoint on the "unprocessed" data and establishes the basis for standardized QA/QC protocols that are consistent across institutions, experts, and over time.

Conclusion

Our report showed that defacing influences human quality assessments of T1w images, with lower-quality nondefaced images likely rated higher after defacing. *MRIQC*'s automatically-extracted image quality metrics (IQMs), on the other hand, remained largely unaffected, as the tool does not consider areas typically altered by defacing when computing the IQMs. We show how privacy protection measures, like defacing, can degrade data reusability and emphasize that actions early in the computational pipeline may have significant downstream effects. We argue that sharing QA/QC materials from nondefaced images can help mitigate these issues while ensuring participant privacy.

Data and code availability statement

All the new materials relating to this work were shared under suitable open licenses (Apache 2.0 for code and CC-BY for data, unless otherwise specified) before submission of the Stage 2 report. The IXI dataset is available at https://brain-development.org/ixi-dataset/ under the Creative Commons CC BY-SA 3.0 license. All the code related to this study and tabular data collected (manual ratings and the related covariates, as well as the IQMs extracted from defaced and nondefaced) are openly shared under the Apache 2.0 license at https://github.com/TheAxonLab/defacing-and-qc-analysis. The Web Service we implemented to collect the manual ratings in this study is available under the Apache 2.0 license at https://github.com/nipreps/qkay. The presentation used in the training session can be accessed at https://osf.io/d57mb/, and the training reports are openly shared at https://github.com/TheAxonLab/defacing-and-qc-trainingsession. The MRIQC reports generated from the IXI dataset, and used by the human raters in their assessment, are accessible via https://github.com/TheAxonLab/defacing-and-qc-ixi-reports under the terms of the CC-BY-SA-4.0 license, as they are derived from the original IXI dataset. All the code related to this study, namely the code to process the IXI dataset and carry out the analyses, is openly shared under the Apache 2.0 license at https://github.com/TheAxonLab/defacing-and-qc-analysis.
Should any of the participants' data be recalled from the original IXI dataset, e.g., after a UK GDPR request, we will accordingly recall the corresponding data derived from the participant's T1w image.

Author contributions

Conceptualization, funding acquisition, project administration, and supervision: OE. Data curation, visualization, and writing: CP and OE. Data collection: CP, TS, EM, JB, EFG, and OE. Methodology: CP, OE, ES, YA, RP, and JR. Formal Analysis: CP, and OE. Software: ES, CP, and OE. Investigation: CP, ES, TS, JB, EM, EGF, and OE. Resources: PH, RP, and OE. All authors reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

This work has been supported by the Swiss National Science Foundation —SNSF— (#185872, OE), the NIMH (RF1MH121867; OE, RP), and the Chan-Zuckerberg Initiative (EOSS5-266; OE). PH, and YAG receive support from SNSF (#185897, PH).

References

1. Schwarz CG, Kremers WK, Wiste HJ, Gunter JL, Vemuri P, Spychalla AJ, et al. Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. NeuroImage. 2021 mai;231:117845.

2. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). OJ L. 2016 May 4;Official Journal of the European Union:1–88.

3. Health Insurance Portability and Accountability Act of 1996 [Internet]. Public Law 104-191 Aug 21, 1996. Available from: https://www.hhs.gov/hipaa/index.html

4. de Sitter A, Visser M, Brouwer I, Cover KS, van Schijndel RA, Eijgelaar RS, et al. Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. Eur Radiol. 2020 Feb 1;30(2):1062–74.

5. Bhalerao GV, Parekh P, Saini J, Venkatasubramanian G, John JP, Viswanath B, et al. Systematic evaluation of the impact of defacing on quality and volumetric assessments on T1-weighted MR-images. J Neuroradiol. 2022 May 1;49(3):250–7.

6. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PLOS ONE. 2017 Sep;12(9):e0184661–e0184661.

7. Esteban O, Ciric R, Finc K, Blair RW, Markiewicz CJ, Moodie CA, et al. Analysis of task-based functional MRI data preprocessed with fMRIPrep. Nat Protoc. 2020 Jul;15(7):2186–202.

8. 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 Feb;59(3):2142–54.

9. Zalesky A, Fornito A, Cocchi L, Gollo LL, van den Heuvel MP, Breakspear M. Connectome sensitivity or specificity: which is more important? NeuroImage. 2016 Nov;142:407–20.

10. 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. Hum Brain Mapp. 2016 Jul;37(7):2385–97.

11. Esteban O, Poldrack RA, Gorgolewski KJ. Improving Out-of-Sample Prediction of Quality of MRIQC. In: Stoyanov D, Taylor Z, Balocco S, Sznitman R, Martel A, Maier-Hein L, et al., editors. Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Cham: Springer International Publishing; 2018. p. 190–9. (Lecture Notes in Computer Science).

12. Keshavan A, Datta E, M. McDonough I, Madan CR, Jordan K, Henry RG. Mindcontrol: A web application for brain segmentation quality control. NeuroImage. 2018 Apr 15;170:365–72.

13. Keshavan A, Yeatman JD, Rokem A. Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. Front Neuroinformatics [Internet]. 2019 [cited 2022 Apr 22];13. Available from: https://www.frontiersin.org/article/10.3389/fninf.2019.00029

14. Marcus DS, Harms MP, Snyder AZ, Jenkinson M, Wilson JA, Glasser MF, et al. Human Connectome Project informatics: Quality control, database services, and data visualization. NeuroImage. 2013 Oct 15;80:202–19.

15. Shezhad 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. Front Neurosci [Internet]. 2015 [cited 2022 Jul 25];9. Available from: http://www.frontiersin.org/Community/AbstractDetails.aspx?ABS_DOI=10.3389/conf.fnins.2015.91.00047

16. Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage. 2018 Feb 1;166:400–24.

17. Shehzad Z, Steven G, Qingyang L, Yassine B, Chaogan Y, Zhen Y, et al. The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of

MRI data. Front Neurosci [Internet]. 2015 [cited 2022 Apr 22];9. Available from: http://www.frontiersin.org/Community/AbstractDetails.aspx?ABS_DOI=10.3389/conf.fnins.2015.91.00047

18. Mortamet B, Bernstein MA, Jack CR, Gunter JL, Ward C, Britson PJ, et al. Automatic quality assessment in structural brain magnetic resonance imaging. Magn Reson Med. 2009 Aug;62(2):365–72.

19. Garcia M, Dosenbach N, Kelly C. BrainQCNet: a Deep Learning attention-based model for the automated detection of artifacts in brain structural MRI scans. Imaging Neurosci [Internet]. 2024 Sep 11 [cited 2024 Sep 17]; Available from: https://doi.org/10.1162/imag_a_00300

20. Provins C, Alemán-Gómez Y, Cleusix M, Jenni R, Richiardi J, Hagmann P, et al. Defacing biases manual and automated quality assessments of structural MRI with MRIQC [Internet]. Glasgow: 28th Annual Meeting of the Organization for Human Brain Mapping (OHBM) (Glasgow). p. WTh566; 2022 [cited 2022 Apr 4]. Available from: https://osf.io/t9ehk

21. Hill D, Williams S, Hawkes D, Smith S. IXI Dataset – Brain Development [Internet]. 2006 [cited 2022 Feb 18]. Available from: https://brain-development.org/ixi-dataset/

22. Gulban OF, Nielson D, Poldrack R, Lee J, Gorgolewski C, Sochat V, et al. poldracklab/pydeface: v2.0.0 [Internet]. Zenodo; 2019 [cited 2021 Dec 7]. Available from: https://zenodo.org/record/3524401

23. Theyers AE, Zamyadi M, O'Reilly M, Bartha R, Symons S, MacQueen GM, et al. Multisite Comparison of MRI Defacing Software Across Multiple Cohorts. Front Psychiatry. 2021;12:189.

24. Provins C, MacNicol E, Seeley SH, Hagmann P, Esteban O. Quality control in functional MRI studies with MRIQC and fMRIPrep. Front Neuroimaging [Internet]. 2023 [cited 2023 Jan 12];1. Available from: https://www.frontiersin.org/articles/10.3389/fnimg.2022.1073734

25. Savary E, Provins C, Sanchez T, Esteban O. Q'kay: a manager for the quality assessment of large neuroimaging studies. In 29th Annual Meeting of the Organization for Human Brain Mapping (OHBM); 2023 [cited 2023 Feb 20]. Available from: https://osf.io/edx6t/

26. Nárai Á, Hermann P, Auer T, Kemenczky P, Szalma J, Homolya I, et al. Movement-related artefacts (MR-ART) dataset of matched motion-corrupted and clean structural MRI brain scans. Sci Data. 2022 Oct 17;9(1):630.

27. Taylor PA, Glen DR, Reynolds RC, Basavaraj A, Moraczewski D, Etzel JA. Editorial: Demonstrating quality control (QC) procedures in fMRI. Front Neurosci [Internet]. 2023 [cited 2023 Oct 3];17. Available from: https://doi.org/10.3389/fnins.2023.1205928

28. 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. Mol Psychiatry. 2014 Jun;19(6):659–67.

29. Gorgolewski KJ, Durnez J, Poldrack RA. Preprocessed Consortium for Neuropsychiatric Phenomics dataset [Internet]. F1000Research; 2017 [cited 2021 Oct 28]. Available from: https://f1000research.com/articles/6-1262

30. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples)†. Biometrika. 1965 Dec 1;52(3–4):591–611.

31. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots [Internet]. 2020 [cited 2022 Sep 12]. Available from: https://CRAN.R-project.org/package=ggpubr

32. R Core Team. R: The R Project for Statistical Computing [Internet]. 2021 [cited 2024 Sep 6]. Available from: https://www.r-project.org/

33. Mauchly JW. Significance Test for Sphericity of a Normal n-Variate Distribution. Ann Math Stat. 1940 Jun;11(2):204–9.

34. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests [Internet]. 2023 [cited 2022 Aug 2]. Available from: https://CRAN.R-project.org/package=rstatix

35. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009 Nov;41(4):1149–60.

36. Bates D, Maechler M, Bolker [aut B, cre, Walker S, Christensen RHB, et al. lme4: Linear Mixed-Effects Models using "Eigen" and S4 [Internet]. 2022 [cited 2022 Sep 12]. Available from: https://CRAN.R-project.org/package=lme4

37. Kirk, Roger E. 2012. "Experimental Design: Procedures for the Behavioral Sciences. " SAGE Publications.

38. Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. J R Stat Soc Ser Stat. 1983;32(3):307–17.

39. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999 Apr 1;8(2):135–60.

40. Friedrich S, Konietschke F, Pauly M. Resampling-Based Analysis of Multivariate Data and Repeated Measures Designs with the R Package MANOVA.RM. R J. 2019;11(2):380–400.

41. Morgan CA, Roberts RP, Chaffey T, Tahara-Eckl L, van der Meer M, Günther M, et al. Reproducibility and repeatability of magnetic resonance imaging in dementia. Phys Med. 2022 Sep 1;101:8–17.

42. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

43. Li Q, Zhang J, Dai S. On estimating the non-centrality parameter of a chi-squared distribution. Stat Probab Lett. 2009 Jan 1;79:98–104.

44. Gudicha DW, Schmittmann VD, Vermunt JK. Power Computation for Likelihood Ratio Tests for the Transition Parameters in Latent Markov Models. Struct Equ Model Multidiscip J. 2016 Mar 3;23(2):234–45.

45. Wilk MB, Gnanadesikan R. Probability Plotting Methods for the Analysis of Data. Biometrika. 1968;55(1):1–17.

46. Hoenig JM, Heisey DM. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. Am Stat. 2001 Feb 1;55(1):19–24.

47. Ben-Shachar MS, Lüdecke D, Makowski D. effectsize: Estimation of Effect Size Indices and Standardized Parameters. J Open Source Softw. 2020 Dec 23;5(56):2815.

48. 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. Front Neuroinformatics [Internet]. 2016 Dec 19 [cited 2024 Aug 8];10. Available from: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2016.00052/full