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 step 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 modeling their ratings on the same set of images in two conditions: "nondefaced" (i.e., 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 (GDPR) in Europe [2] or the Health Insurance Portability and Accountability Act (HIPAA) in the US [3]. 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] investigated the undesired effects of defacing on

automated workflows in neurodegeneration studies. Failure rates—i.e., the percentage of inputs for which workflows did not finish successfully—were substantially larger on defaced inputs (19%) compared to their nondefaced counterparts (2%). They also reported systematic differences between the same processing with and without defacing in several outcomes of interest, such as gray- and white-matter volume and cortical thickness. 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.

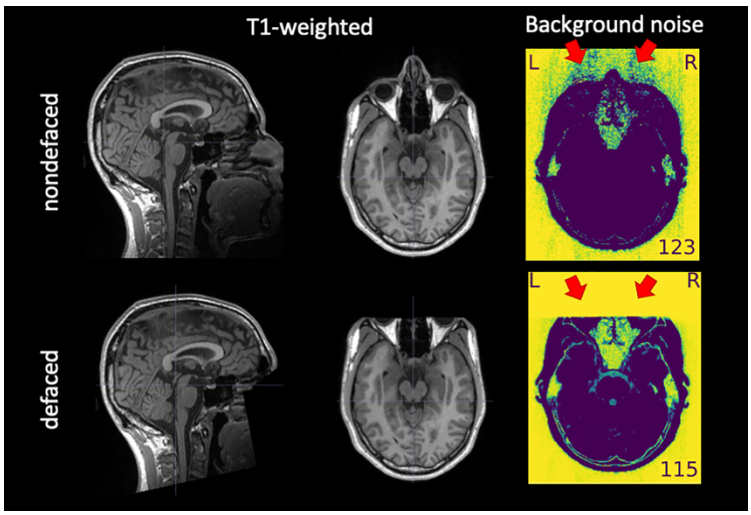

Figure 1. An example of T1w image before and after defacing. Defacing is typically implemented by zeroing the voxels around the face (left-hand side, panel "T1-weighted"). The "Background noise" panel shows two visualizations extracted from the *MRIQC* visual report, in which a window is applied to select the lowest intensities and then they are inverted to enhance patterns in the brackground. The red arrows indicate aliases along the anterior-posterior axis—which had the lowest bandwidth in the example—produced by eye motion. These aliases are straightforward to notice in front of the ocular globes in the "Background noise" panel of the "nondefaced" image because of the absence of signal sources. This aliasing also spreads in the opposite direction, overlapping brain tissue. However, this overlap is often very hard to notice against the signal of interest within the brain. The corresponding "defaced" version of the "Background noise" panel shows how defacing eliminates valuable information for quality assessment.

Understanding how defacing alters the initial quality 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). Our previous efforts to automate QC by learning *MRIQC*'s IQMs [6, 11] demonstrated that site-effects imposed the largest limitation on the performance. The reportedly modest reliability of automated QC alternatives currently available explains that QA/QC

continues to require the screening of 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 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 quality assurance (QA) protocols and corresponding QC criteria, e.g., the Human Connectome Project [14] or the INDI initiative (QAP [15]). One related but 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 manual 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 [6,15,17] or by training deep models directly on 3D images [18]. 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 [19], 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 built on our exploratory analysis and confirmed that four trained human raters made systematically varying QC decisions when rating the same images depending on whether they have been defaced. Regarding automated QC, this paper showed that differences in IQMs extracted with *MRIQC* before and after defacing were not meaningful.

Methods

The lettered footnotes in this section indicate deviations from the Stage 1 protocol and are reported in the “Protocol Deviations” subsection.

Hypotheses

This pre-registered report was set out to confirm whether defacing alters the manual and automatic QC 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 [19], 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 analysis was based on the publicly available IXI dataset [20], which contains 581[a] 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 scan acquired at the 3T site (HH)", yielding a total of 185 subjects[b]. In testing Hypothesis 2, we included all participants in the IXI dataset with a T1w scan (581[a] subjects)[b]. Images were excluded from the evaluation of hypotheses 1 and 2 in the case of complete failure of image reconstruction[b]. The local ethics committee has approved the processing of nondefaced images. This study does not attempt to re-identify the participants nor facilitate in any way such efforts[c].

Data processing. First, a defaced version of each scan (N=581) was generated with *PyDeface* [21]. We chose *PyDeface* because it currently presents the highest success rate at removing facial features while not removing brain voxels [22]. 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 exclusion criterion. Because all images were successfully processed, no images were dropped from the analysis for this reason. To avoid exclusion bias that could artificially inflate the effect size, images showcasing low-quality defacing, such as those partially retaining facial features—e.g., the eyes,—were included in the analysis[d]. These cases were expected to be more similar between defaced and nondefaced conditions, potentially attenuating differences in human ratings and IQMs. The raters then assessed the quality of the same images in two conditions (nondefaced and defaced) following the Manual assessment protocol described below. Raters did not have access to the mapping between defaced and nondefaced counterparts. We obfuscated participant identifiers and shuffled their ordering before presentation. 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[e].

Manual assessment protocol. We performed manual quality assessment only on the images from the site with a 3T device (HH; N=185). 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 [f]) assessed the quality of the subsample in each of the two conditions (that is, defaced and nondefaced). The quality assessment was carried out by individually screening one *MRIQC*-generated visual report per subject and condition. These reports are openly shared (see Data and code availability statement). Raters were recruited by inviting researchers who expressed their interest in QA/QC within the Department of Radiology of the Lausanne University Hospital (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 the consistency of their training, raters read our published QC protocol [23] and participated in a 3h [g] training session. Following our guidelines [23], raters were instructed that it was critical to scrutinize the background visualization carefully because most of the artifacts are more noticeable in the absence of signal of interest. Several examples supported this remark during the training sessions (see Fig. 1, Supplementary Fig. S21, and the training materials). At the end [g] of this session, raters self-assessed their experience as either beginner, intermediate, or advanced. [h] One rater identified themself as an advanced rater, the three others as intermediate. The training session included an introduction to *MRIQC*'s visual reports, its "rating widget" (Fig. S1), and the data collection tools and settings. We also presented the list of quality criteria to consider, and for each criterion, we illustrated an exclusion example. These examples were extracted from diverse, publicly available datasets [24–27] and in-house datasets. Lastly, in the same session, the raters were asked to assess 20 images independently, and we compared and discussed the ratings together to ensure that the quality criteria were interpreted consistently. None of the training and calibration examples were extracted from the IXI dataset. 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 QC, 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. 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 then 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 presented in Fig. S1 [h]. A quality score was assigned using a

slider that permits the selection of numbers in 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 [h], 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 [23], 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 recorded the exact time each assessment took. The time for each assessment was measured and is available for future exploration. The removal of the IQMs from the visual reports, the assignment of images in the two conditions to raters, the shuffling of ordering of presentation, the actual presentation, and the tracking of raters' progress was managed with *Q'kay* [28], a Web Service we developed for this study. [b] Images assigned the lowest grade (one in our 1-4 interval scale) in both conditions by all raters would be excluded from all analyses. No images met this 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) by comparing linear mixed-effects (LME) model fits by means of a likelihood-ratio test. The LME modeling was a contingency alternative in case data violated some of the normality and sphericity assumptions of repeated-measures ANOVA (rm-ANOVA). As opposed to multiple *t*-tests, rm-ANOVA and LMEs enabled disentangling the inter-rater variability from intra-rater variability due to defacing and quantifying the latter. Because we do 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. We removed the 40 repeated images from the dataset because neither rm-ANOVA nor LMEs allow non-uniquely identified rows [i]. We first verified that the sphericity and normality assumptions of rm-ANOVA were unmet. Normality was verified with the Shapiro-Wilk normality test [29], employing the *shapiro.test* function of the *ggpubr R* package [30]. Sphericity was assessed with Mauchly's test for sphericity [31] within the *rstatix* [32] *R* package. rm-ANOVA was implemented with the *anova_test* function (*rstatix*; $p<.02$). [i] A preliminary sensitivity analysis (Table 1) executed with *G*Power* [33] had determined 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% (see Equation S1 to convert effect size of type f to type η_p^2). For context, we found an effect size of f=0.31 (see Equations S2 and S3) in our pilot study. Comparisons between both effect sizes need to be performed cautiously as the design of the rating collection has been modified between the pilot study and this report. 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., subject and rater were random effects) [j]. The models were implemented in R with the *lmer* function of the *lme4* package [34]. As part

of regression diagnostics (e.g., non-Gaussian or heteroscedastic residuals indicate non-optimal model fit), we examined the shape of 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* [35] [k]. We set the significance threshold for the likelihood ratio at p=0.02. In addition, to estimate the size of the effect, we computed the noncentrality parameter λ associated with the likelihood ratio test, which is a proxy for its power [36]. This post-hoc power analysis was then compared to the minimum power achievable from a pre-registered sensitivity analysis, which indicated that $\lambda \geq 13.017$ would allow the detection of differences between models at 90% power (see Table 1) [l]. 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* [33]. This sensitivity corresponds to η_p^2 =0.045 (i.e., a medium effect size) following Equation S1. 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. S11), 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		

Confirming that—on average—ratings are higher on defaced images. We used Bland-Altman (BA) plots [37] to visualize the bias and the 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 changes depending on the rater. [m]The bias $\mathbb{E}[\Delta_{\text{ndef}-\text{def}}]$ is defined [37] as the mean of rating differences $\Delta_{\text{ndef}-\text{def}} = r^s_{\text{ndef}} - r^s_{\text{def}}$, where $r^s_{\text{condition}}$ are the ratings corresponding to the nondefaced (ndef) and defaced (def) conditions of a single subject's image s. Therefore, to investigate our hypothesis that humans' *ratings are higher on defaced images*, we confirmed that $\mathbb{E}[\Delta_{\text{ndef}-\text{def}}] < 0$ for all raters, aggregated and individually (Fig. 2). Bland & Altman [37] recommend assessing the agreement between measurements by calculating the 95% limits of agreement (LoA), defined as *the range in which 95% of the differences* $\Delta_{\text{ndef}-\text{def}}$ *are expected to fall.* In practice, *two measurement approaches cannot be used interchangeably if their LoA are too broad.* Following [37], LoA are calculated as follows:

$$\text{LoA} = \left(\mathbb{E}[\Delta_{\text{ndef}-\text{def}}] - z_{\alpha/2} \cdot \sigma,\ \mathbb{E}[\Delta_{\text{ndef}-\text{def}}] + z_{\alpha/2} \cdot \sigma\right),$$

where σ is the sample standard deviation, and $z_{\alpha/2}$ is the critical value—with $z_{\alpha/2}$=1.96 for the 95% coverage range of the differences, and $z_{\alpha/2}$=2.0 recommended as a close approximation [37]. To demonstrate that nonzero biases were not obtained by chance, we tested whether the bias was significantly negative by checking if $\Delta_{\text{ndef}-\text{def}}$=0 was contained within the 95% confidence interval (CI) for the bias [37]. We calculated the 95% CI for the bias under the assumption of normality and reasonably large sample sizes following Bland & Altman's indications [37], with:

$$\text{CI} = \left(\mathbb{E}[\Delta_{\text{ndef}-\text{def}}] - 1.96 \cdot \frac{\sigma}{\sqrt{n}},\ \mathbb{E}[\Delta_{\text{ndef}-\text{def}}] + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right),$$

where n is the sample size. To protect from violations of the normality assumption, we also computed them via bootstrapping [38]. We reported both estimates and tested for the significance with the more conservative option. Additional considerations on the interpretation of BA plots, the 95% LoA, and the 95% CI of the mean bias are provided in the Supplementary Materials, subsection S6.2.

Determining that defacing introduces biases in *MRIQC*-generated IQMs. Defacing impact on automatic QC was evaluated based on 58 [n] out of 62 IQMs calculated by *MRIQC*. For the complete list of IQMs produced by *MRIQC* and their definitions, refer to Table 2 in [6]. A two-way repeated-measures, multivariate ANOVA (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 [39] in R. Because many IQMs were heavily correlated (see Fig. S14), their dimensionality was reduced with principal components analysis (PCA) before applying rm-MANOVA. To ensure the defacing effects were not mitigated, PCA was applied only on the IQMs extracted from the nondefaced data, and the resulting transformation was applied to IQMs to the entire dataset. PCA was implemented with the *prcomp* function of R's *stats* package, 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 [11,40]. We applied a significance level of $p<0.02$ for the rm-MANOVA and considered the p-values extracted under the Wald-type statistics section. A preliminary sensitivity analysis determined that our experimental design can 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* (see Table 2). For context, the effect size associated with the MANOVA on the IQMs of our pilot study was f=0.16 (see Equations S4 and S5 for its computation). However, comparisons between both effect sizes need to be exercised with caution as the statistical designs are different. In our pilot study, we used a standard MANOVA on only five IQMs that showed the strongest bias on the BA plots. 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 reported in [5] for the *PyDeface*'s influence on IQMs ranged from f=0.045 to f=1.79, with a mean effect size across IQMs of f=0.61 (see Equation S3 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. S15 through S18 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	Noncentrality 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

Protocol Deviations

[a] **Erroneous dataset size reported within the pre-registered plan.** The IXI dataset contains 581 images, but we erroneously stated that it had 580 T1w images in the pre-registration. This error only affected this experiment because the reported number of participants in the HH subsample for Hypothesis 1 (N=185) was correctly reported.

[b] **Improved clarity in the description of inclusion and exclusion criteria from the IXI dataset.** Our approved Stage 1 registration left much room for precision on the inclusion and exclusion criteria. The final text clearly delineates these criteria, separated by the two hypotheses. It also resolves the problem that while we originally had proposed three hypotheses, the finally accepted Stage 1 had

been reduced to the two we introduce here. Unfortunately, some text referred to hypothesis 3, which had been dismissed after peer review. This deviation also changes the location of some exclusion criteria, for example, those regarding later steps of the processing stream that are now placed at the corresponding point of the methodological description (e.g., excluding images due to insufficient quality of defacing). While the exclusion/inclusion criterion did not change, we now indicate the criterion later in the protocol.

[c] **Moved ethical statement and GDPR enforcement to more appropriate locations.** The ethical statement has been moved up to the initial paragraph of the "Data" section. The GDPR enforcement statement is now with the "Data and software availability statement".

[d] **Improved precision of the QC checkpoint after *PyDeface*.** Our approved Stage 1 should have been more specific regarding what "ineffective" defacing means. Fortunately, this exclusion criteria did not apply, and no images were excluded for this reason.

[e] **Statement of exclusions due to *MRIQC* failures.** The approved Stage 1 plan did not explicitly anticipate the potential failure of *MRIQC* when processing images. We report that *MRIQC* unsuccessfully processed two subjects.

[f] Expert raters were incorporated as authors after data collection.

[g] The training session lasted 3 hours instead of 4 hours. Raters self-assessed their experience at the end of the session rather than at the beginning.

[h] The training session has been comprehensively reported, and the protocol updated correspondingly. Figure 2 in the pre-registered plan has been moved to the supplementary materials and is now Fig. S1.

[i] **Images rated twice for the intra-rater reliability assessment could not be fed into the planned analyses.** We had not tested at pre-registration time that neither the *rstatix*'s implementation of rm-ANOVA nor the LMEs' implementation in *lme4* could deal with non-uniquely identified rows. As such, in the case of the images repeated twice, we excluded the ratings corresponding to the second image presentation for all raters in both conditions. We updated the preliminary sensitivity analysis for the rm-ANOVA modeling according to the reduced sample, changing the smallest detectable effect size at 90% power from f=0.14 to f=0.218. We have also transformed the reporting of the sensitivity analysis from the screen capture of *G*Power* in the former Figure 3 into the new Table 1. This change had no notable consequences as we finally carried out modeling with the LMEs alternative.

[j] **Subject as a random effect in the LMEs.** Our pre-registered plan failed to identify the subject as a random effect in the LMEs, which is critical for the model to capture that the subject is a repeated measure.

[k] **Added a reference**. We added the citation to the R software package [35] that was missing in the pre-registration.

[l] **Post-hoc power analysis.** Our pre-registered plan stated that "We will deem the effect irrelevant if the latter parameter is smaller than 13", which we later understood is a misinterpretation of the utility of post-hoc power analyses and discussed in the results.

[m] **Corrected interpretation of BA plots.** Our pre-registered plan employed 95% LoA of the difference and 95% CI of the bias as interchangeable terms, which is incorrect. Our Stage 2 report was updated to employ unambiguous nomenclature for the bias $\mathbb{E}[\Delta_{\mathrm{ndef-def}}]$, included the definition of LoA and CI as proposed initially by Bland & Altman [37, 38], and provided further details regarding the interpretation of LoA and the significance of the biases.

[n] Two IQMs, Mortamet [17] "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. S12, with the plots corresponding to the four excluded IQMs highlighted in red.

Methodological precisions at Stage 2 and exploratory analyses

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.

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 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 1, 2, and 3 because of the stark difference in Rater 4'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 six tests controlling for false discovery rate (FDR [41]).

Post-hoc sensitivity analysis implementation. Our pre-registered plan committed to estimating the effect size of the likelihood-ratio test without methodological details. 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 [42,43]. 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 [42] 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 explored 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. S13. 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 [44] 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 false discovery rate (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 [29] (see Table S3; Fig. S9). Nonetheless, we also carried out the planned rm-ANOVA for completeness as an exploratory analysis, 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, 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" (see deviation [l]), as effect size estimates can be extremely biased [45]. 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 units. 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). Although all raters displayed individual LoA larger than one unit, indicating that defacing introduced large variability, these plots also reflected that raters were not equally affected by defacing. Our most experienced rater (Rater 1) presented the highest bias $\mathbb{E}[\Delta_{\mathrm{ndef-def}}]$ = -0.12 (Table S2), while a more junior rater (Rater 4) barely modified their quality assessment ($\mathbb{E}[\Delta_{\mathrm{ndef-def}}]$= -0.02). The differences between nondefaced and defaced ratings by Rater 4 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 4 supported that Raters 1, 2, and 3 showed better agreement, as shown by the larger rater effect size compared to our pre-registered model counterparts with all ratings (Table S4). $\mathbb{E}[\Delta_{\mathrm{ndef-def}}]$ was consistently negative across all raters, implying that defaced images were, on average, scored with a higher rating. This bias was significant for Raters 1 and 3, and all raters aggregated, as shown in Fig. 2 and comprehensively reported in Table S2.

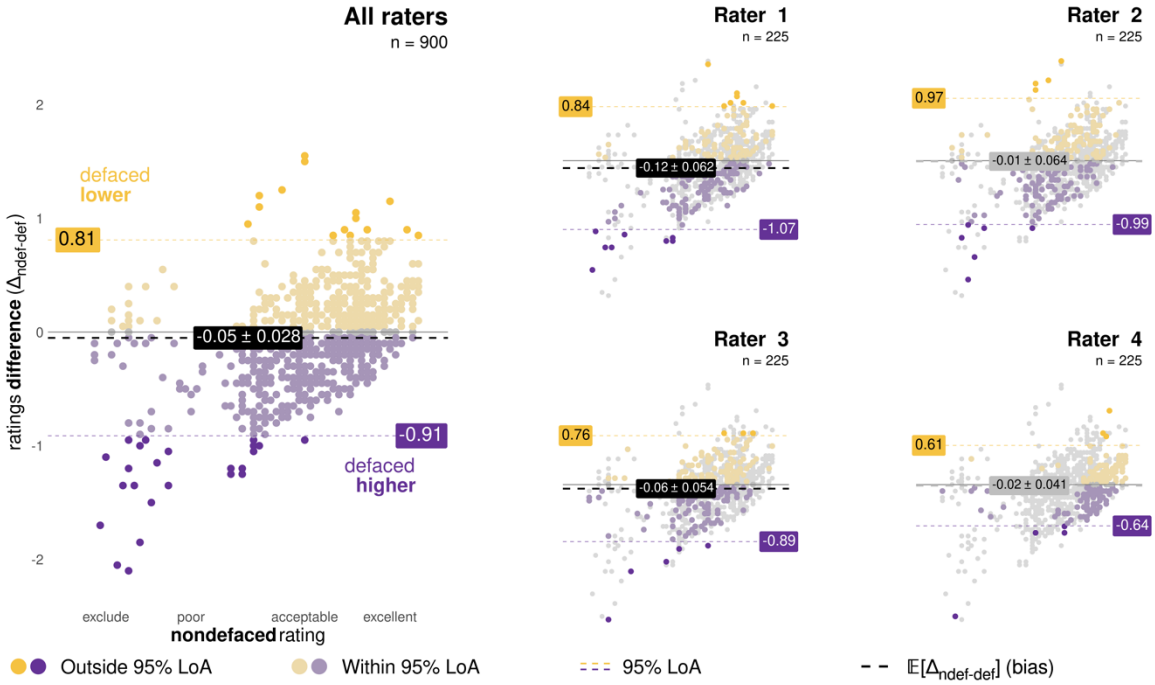

Figure 2. Defacing biases human assessment of image quality, particularly when image quality is low. We examined biases with an "optimized" version of the BA plot, in which the *x*-axis represents the rating assigned to the nondefaced version of an image. Corresponding "standard" BA plots—in which the *x*-axis shows the average of the two ratings [37]—are reported in Fig. S2. Rating pairs where the defaced image's quality was underestimated with respect to the nondefaced ($\Delta_{\mathrm{ndef-def}} > 0$) are represented in yellow. Conversely, pairs where the defaced image's quality was overestimated ($\Delta_{\mathrm{ndef-def}} < 0$) are in purple. Pairs within the 95% LoA are represented with dim colors, and the LoA boundaries are indicated with dashed colored lines annotated with their value. For example, the LoA for all raters pooled together was [-0.91, 0.81] (left panel). Finally, the bias $\mathbb{E}[\Delta_{\mathrm{ndef-def}}]$ is represented by a gray or black dashed line with a label reporting value and their corresponding 95% CI interval (parametric estimation). All raters displayed 95% LoA exceeding one unit, indicating that defacing introduces large variability in human assessments. All raters had negative—albeit small—biases, indicating that they systematically rated defaced images higher. However, these biases were statistically significant only for Raters 1 and 2, as well as the four raters aggregated together—indicated by the bias label and line colored in black. Relevant statistics (bias, LoA, 95% CI) are reported in Table S2, and the full report of statistics, including 95% CI intervals calculated both by parametric and non-parametric means are distributed as tab-separated-values files on the corresponding GitHub repository.

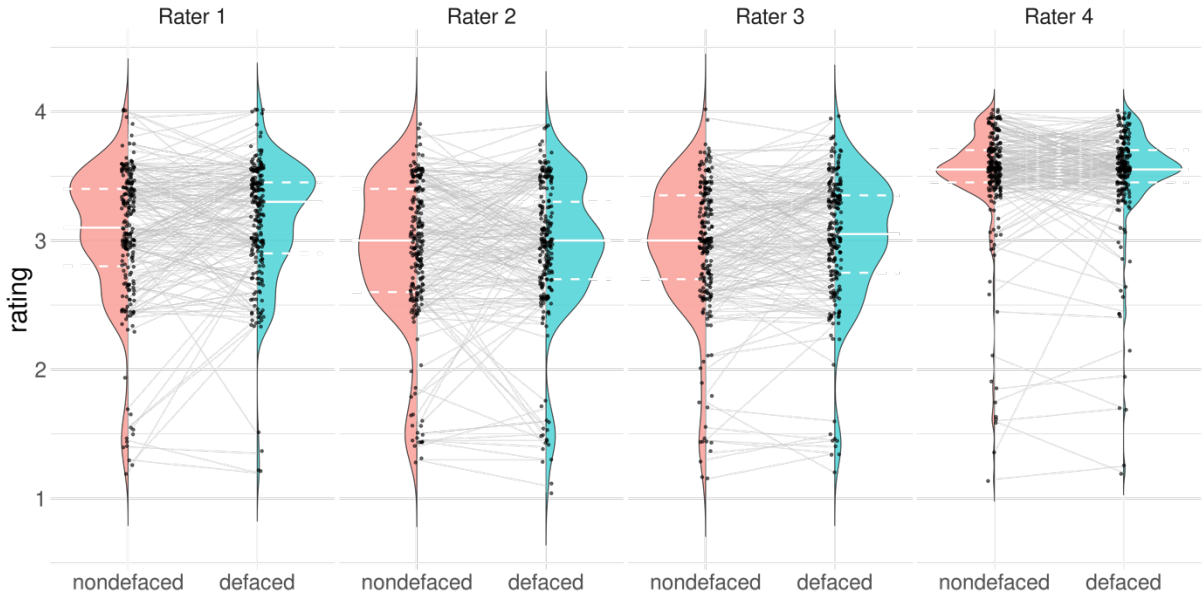

Figure 3. Rater 4 issued visibly different ratings, generally more optimistic, than the other raters in both (defaced and nondefaced) conditions. The gray lines highlight the evolution of the rating between the nondefaced image and its defaced counterpart. The full white line in the violin plot represents 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 1) showed the largest bias. Rater 4's rating distribution diverged from that of the other raters, being more optimistic overall about the quality of the images. Rater 4 also displayed a lower spread in quality assessments, which translated into the narrowest 95% LoA (Fig. 2). 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. 2 and S2).

***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, $\eta_p^2 = 3.5 \cdot 10^{-6}$) nor between sites (F(2, 184) = 0.376, p=0.828, $\eta_p^2 = 7.9 \cdot 10^{-4}$). As explained in the section "Methodological precisions at Stage 2 and exploratory analyses", 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. S19) 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< 0.001, η_p^2=0.0047, f=0.069; see Equation S1). 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 and 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. S15), we found that nearly all IQMs showcased very narrow 95% LoA and minimal biases—yet non-zero—compared to their variation ranges. The narrow LoA indicate that the variance introduced by defacing into the IQMs is relatively small. Biases were deemed significant in most of the IQMs, which immediately follows the relatively large sample size. Only the entropy-focus criterion

("efc") emerged as an exception, showcasing a substantial bias and a broad 95% LoA, which, in addition, did not contain zero. As opposed to all other IQMs, the bias corresponding to the "efc" was also clearly noticeable in violin plots (Fig. S13).

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. 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* [46], which implements the formula given in Equation S4 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

Typical neuroimaging workflows involve many image-processing steps, contributing to analytical variability that cascades throughout the pipeline. Even initial processing steps considered relatively innocuous, such as reorienting the volume to align the anterior-posterior commissure line with the axial plane, may conceal relevant design decisions and inadvertently introduce variability. Because defacing has been shown to influence downstream processing [1, 4-5], this registered report investigated the alterations introduced by this step in the manual and automatic QC of whole-brain T1w 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 $\mathbb{E}[\Delta_{\mathrm{ndef-def}}]$ = -0.12 (Table S2) was displayed by the most experienced rater (Rater 1). This indicates that human raters are sensitive to defacing, possibly more for more experienced raters. The first implication of this result is that QA/QC protocols establishing a visual checkpoint on the unprocessed data [23]—e.g., with *MRIQC*'s visual reports—should generate visualization and other QA/QC materials before defacing anatomical images. Defacing is generally required for compliance with applicable

privacy protection regulations if data are to be processed on systems with low-security clearances or if data will be made accessible to third parties or publicly. Regardless of the motivation to deface anatomical images, our results support the recommendation of providing QA/QC materials extracted from the original (nondefaced) images alongside the defaced dataset, 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 settings more consistent with those of the original authors, minimizing the risk of artifacts going unnoticed and biasing analyses.

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/QC attempts to minimize data exclusion on the grounds of insufficient quality), and it is more acute in public datasets (because 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 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 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 protocols [23].

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 distributions of raters' assessments diverged 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 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 resources. Nonetheless, it cannot be ruled out that other circumstances

determined the differences in rating distribution, such as attrition drifts (Rater 4 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 assessments sticking out compared to the others, our study was 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, only the "efc" metric showed a meaningful difference between conditions out of the total 58 IQMs investigated. Although most IQMs had significant biases (Fig. S15) due to the large sample size, these biases were minimal, and their LoA were narrow in comparison to their corresponding variation ranges. Conversely, "efc" was the only metric displaying a broad LoA compared to its typical range (Fig. S15). In addition, except for "efc", no other IQM's LoA excluded the zero-difference line, indicating that very rarely, differences in "efc" will be small between conditions. This result seems a consequence of *MRIQC*'s definition of most metrics, such as contrast-to-noise ratio ("cnr") or signal-to-noise ratio in the gray matter ("snr_gm"), which are extracted from areas within the brain and, therefore, unaffected by defacing. The "efc" metric is an exception among IQMs, including background in their definition. For example, the "snrd_*" family of metrics is akin to their "snr_*" counterparts with the difference that the signal variability is estimated in the background [48]. While "efc" employs all the background surrounding the head and neck, "snrd_*" metrics discard background information outside of a "hat" mask that excludes the background below the nasio-occipital plane. Hence, they are also insensitive to the area altered by defacing. Our results contrast with [5], which 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. S13). These reasons explain why they found significant differences with a 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. Prior work by Pizarro et al. [49] on anatomical MRI and our later work on functional MRI [50] justified how information "outside" the brain region is relevant to assess the quality of imaging "within" the region of interest. These metrics—e.g., those proposed in [49] and derivations thereof—should be clearly annotated and excluded from analyses involving defaced images where they are not applicable. 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 assessing anatomical T1w MRI of the human brain. It provides a tooling framework to establish a rigorous QA/QC checkpoint on the unprocessed data. Finally, this report provides the basis for standardized QA/QC protocols that are consistent across institutions, experts, and over time.

Conclusion

The initial processing steps in the computational pipeline may have relevant downstream effects. Our report showed that defacing influences human quality assessments of T1w images, with lower-quality nondefaced images likely rated higher after defacing. Inadequate manual QC decisions deriving from the visualization of defaced images could undermine the soundness of results. *MRIQC*'s automatically extracted IQMs, on the other hand, remained largely unaffected, as the tool does not consider areas typically altered by defacing when computing the IQMs.

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 the Stage 2 report submission. 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.

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), EFG receives support from SNSF (#10000706, EFG).

Financial Disclosure Statement

The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Supplementary Materials' list of legends

S1 Table	The IXI dataset was acquired at three different hospitals in London (UK).
S2 Table	Defacing-derived bias was larger and more consistent than intrinsic intra-rater variability.
S3 Table	Normality tests.
S4 Table	Results of repeated-measures ANOVA.
S5 Table	Results of running one repeated-measures ANOVA for each rater separately.
S6 Table	Model fit indices of LMEs with ("alternative") and without ("baseline") defacing as a fixed effect.
S7 Table	Results of likelihood ratio test comparing LMEs with and without defacing as a fixed effect.
S8 Table	Variance of random effects coefficients and residuals.
S9 Table	Noncentrality parameter.
S10 Table	Eigenvalues associated with the principal components.
S11 Table	Results of repeated-measures MANOVA on the projected IQMs.
S1 Figure	The updated MRIQC rating widget to assign quality grades using a slider.
S2 Figure	Optimizing BA plots to overcome the wrapping effect of our quality scale.
S3 Figure	Exploring the variability introduced at the tails of the nondefaced ratings distribution.
S4 Figure	Exploring the correlation between nondefaced and defaced ratings.
S5 Figure	Ratings' evolution throughout different rating sessions.
S6 Figure	Test-retest violin plots comparing the rating assigned to the first and second presentation of the repeated images.
S7 Figure	The test-retest exploration did not reveal consistent intra-rater biases.
S8 Figure	Optimized BA plots of intra-rater variability by defacing condition.
S9 Figure	Check for rating normality using a Q-Q plot.
S10 Figure	Diagnostic plots for all LMEs.
S11 Figure	The noncentrality parameter associated with the likelihood ratio test, a proxy for its power (Li, Zhang, and Dai 2009), is 13.017.
S12 Figure	Distribution of the IQMs.
S13 Figure	IQMs distribution per site and condition.
S14 Figure	Several IQMs were highly correlated.
S15 Figure	Bland-Altman plots for all non-excluded IQMs.

S16 Figure	Bland-Altman plots for all principal components extracted with a single standardization and PCA step.
S17 Figure	Bland-Altman plots for all principal components extracted with standardization per site but a single PCA step.
S18 Figure	Bland-Altman plots for all principal components extracted with standardization and PCA per site.
S19 Figure	Scree plot: Variance explained by the principal components.
S20 Figure	Defacing did not influence the IQMs' reliability.
S21 Figure	Screening the background helps detect artifacts within the brain.

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 May 1;231:117845.

2. European Parliament and of the Council. GDPR (General Data Protection Regulation) - Regulation (EU) 2016/679 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 [Internet]. 2016/679 Apr 27, 2016. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj

3. US Health and Human Services Department. HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act. 45 CFR §164.514. 45 CFR §164.514 Aug 21, 1996.

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;12(9):e0184661.

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

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 1;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 15;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 1;37(7):2385–97.

11. Esteban O, Poldrack RA, Gorgolewski KJ. Improving Out-of-Sample Prediction of Quality of MRIQC. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis LABELS 2018, CVII 2018, STENT 2018 [Internet]. Granada, Spain: Springer; 2018. p. 190–9. (Lecture Notes in Computer Science; vol. 11043). Available from: https://doi.org/10.1007/978-3-030-01364-6_21

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 2023 Mar 16];13. Available from: https://www.frontiersin.org/articles/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. Shehzad Z, Giavasis S, Li Q, Benhajali Y, Yan C, Yang Z, et al. The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of MRI data. In: INCF Neuroinformatics [Internet]. Cairns, Australia: Front Neurosci; 2015 [cited 2017 Feb 27]. Available from: http://www.frontiersin.org/10.3389/conf.fnins.2015.91.00047/event_abstract

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. 2017 Oct 24;

17. 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 1;62(2):365–72.

18. 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 19]; Available from: https://doi.org/10.1162/imag_a_00300

19. 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. In: 28th Annual Meeting of the Organization for Human Brain Mapping [Internet]. Glasgow, Scotland: OSF Preprints; 2022 [cited 2022 Jan 30]. p. WTh566. Available from: https://osf.io/8mcyz/

20. Hill D, Williams S, Hawkes D, Smith SM. IXI dataset - Information eXtraction from Images project (EPSRC GR/S21533/02) [Internet]. 2006 [cited 2013 May 7]. Available from: http://www.brain-development.org/

21. Gulban OF, Nielson DM, Lee J, Poldrack RA, Gorgolewski C, Sochat VV, et al. Open Source Software: PyDeface [Internet]. Zenodo; 2022 [cited 2024 Oct 29]. Available from: https://zenodo.org/records/6856482

22. 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 [Internet]. 2021 Feb 24 [cited 2024 Oct 29];12. Available from: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2021.617997/full

23. Provins C, MacNicol EE, Seeley SH, Hagmann P, Esteban O. Quality Control in functional MRI studies with MRIQC and fMRIPrep. Front Neuroimaging. 2023;1:1073734.

24. 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.

25. 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 Sep 8];17. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2023.1205928

26. 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.

27. Gorgolewski KJ, Durnez J, Poldrack RA. Preprocessed Consortium for Neuropsychiatric Phenomics dataset. F1000Research. 2017 Sep 22;6:1262.

28. Savary E, Provins C, Sanchez T, Esteban O. Q'kay: a manager for the quality assessment of large neuroimaging studies. In: Annual Meeting of the Organization for Human Brain Mapping (OHBM) [Internet]. Montréal, QC, Canada; 2023 [cited 2023 Jan 23]. p. 2467. Available from: https://osf.io/edx6t/

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

30. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots [Internet]. 2023 [cited 2024 Oct 29]. Available from: https://cran.r-project.org/web/packages/ggpubr/index.html

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

32. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests [Internet]. 2023 [cited 2024 Oct 29]. Available from: https://cran.r-project.org/web/packages/rstatix/index.html

33. 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 1;41(4):1149–60.

34. Bates D, Maechler M, Bolker [aut B, cre, Walker S, Christensen RHB, et al. lme4: Linear Mixed-Effects Models using “Eigen” and S4 [Internet]. 2024 [cited 2024 Oct 29]. Available from: https://cran.r-project.org/web/packages/lme4/index.html

35. R Core Team. R: The R Project for Statistical Computing [Internet]. [cited 2024 Oct 29]. Available from: https://www.r-project.org/

36. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences [Internet]. SAGE Publications; 2012 [cited 2024 Oct 28]. Available from: https://methods.sagepub.com/book/experimental-design

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

38. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;8(2):135-160.

39. 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.

40. 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.

41. 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.

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

43. Gudicha DW, Schmittmann VD, Vermunt JK. Power computation for likelihood ratio tests for the transition parameters in latent Markov models. Struct Equ Model. 2016;23(2):234–45.

44. Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis for the analysis of data. Biometrika. 1968 Mar 1;55(1):1–17.

45. 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.

46. 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.

47. 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.

48. Dietrich O, Raya JG, Reeder SB, Reiser MF, Schoenberg SO. Measurement of SNRs in MR images: influence of multichannel coils, parallel imaging and reconstruction filters. JMRI. 2007; 26(2):375-385.

49. 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 Neuroinf. 2016;10.

50. Provins C, Markiewicz CJ, Ciric R, Goncalves M, Caballero-Gaudes C, Poldrack RA, Hagmann P, Esteban O. Quality control and nuisance regression of fMRI, looking out where signal should not be found. In: Proc. Intl. Soc. Mag. Reson. Med. 31 (ISMRM). 2022; pp. 2683, London, England, UK.

Table 5. Study design template. This table summarizes the link between the hypotheses, research questions, analysis plans, sensitivity analysis, and prospective interpretation given different outcomes.

Hypothesis	Question	Sampling plan	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes
Defacing influences trained raters' perception of quality	Do the quality ratings from human raters significantly vary between the defaced and the nondefaced conditions?	There is no previous analysis that can inform us on the effect size. For the rationale on how we chose the sample size, refer to the sensitivity analysis in the fifth column.	We will first verify whether the sphericity and normality assumptions of repeated-measures ANOVA (rm-ANOVA) are met. If they are, a rm-ANOVA will then be implemented in R.	The sensitivity analysis, reported in Table 1, indicates that at worst we will be able to confirm differences in manual ratings of f=0.14 corresponding to $\eta^2 = 0.019$ (i.e a medium effect) or greater.	$p<.02$ will indicate significance of the rm-ANOVA, thus confirming that manual quality ratings significantly vary between the defaced and nondefaced conditions. Conversely, we will interpret $p\geq.02$ as a failure to confirm our hypothesis. In any case, the post hoc power achieved and the Cohen's f effect size will be reported. The effect will be deemed irrelevant if the power achieved is lower than 90% or if the Cohen's f effect size is smaller than the minimum detectable effect size we obtained from the sensitivity analysis.
			In the contingency that at least one of the rm-ANOVA assumptions is violated, we will use linear mixed-effects models instead. To test the effect of defacing, we will perform a likelihood-ratio test comparing the models with and without adding the defaced factor as a fixed effect.	The sensitivity analysis for the likelihood ratio test is reported in Table 1.	The bias of defacing on the manual ratings will be deemed significant if the likelihood-ratio test returns $p<.02$. Conversely, we will interpret $p\geq.02$ as a failure to confirm our hypothesis. Furthermore, the effect will be deemed irrelevant if the non-centrality parameter associated with the likelihood ratio test is smaller than 13, corresponding to the minimum power achievable from the sensitivity analysis.

	Are ratings in the defaced condition higher than the correspondi ng ratings on the nondefaced condition ?		We will use Bland-Altman plots (Altman and Bland 1983) to visualize the bias and the limits of agreement of manual quality ratings between the nondefaced and the defaced condition.		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. A bias in the BA plot will be deemed significant if the 95% limits of agreement do not contain the zero difference. In case the 95% limits of agreement do not contain the zero difference, but the bias is positive, we will alternatively conclude that human raters perceive nondefaced images as having better quality overall. Lastly, in case the 95% limits of agreement contains the zero difference, we will conclude that we failed to verify the consistency of defacing bias on manual ratings.
Defacing biases automatic QA/QC of structural MRI with *MRIQC*	Do the IQMs computed by *MRIQC* significantly vary between the defaced and the nondefaced condition ?	As a reference to the sensitivity analysis in the fifth column, the effect size associated with PyDeface influence on IQMs in (Bhalerao et al. 2022) ranged from f=0.045 to f=1.79 with a mean effect size across IQMs of f=0.61.	A two-way repeated-measures MANOVA (rm-MANOVA) will be used to test whether defacing significantly influences the IQMs. However, because many IQMs are heavily correlated (see Figure S14), we will apply principal components analysis (PCA) on the IQMs before running rm-MANOVA.	The sensitivity analysis, reported in Table 2, indicates that we will be able to confirm differences in IQM of f=0.16 corresponding to $\eta^2 = 0.025$ (i.e a medium effect) or greater.	p<.02 will indicate significance of the rm-MANOVA, thus confirming that the IQMs generated by *MRIQC* significantly vary between the defaced and nondefaced conditions. We will consider the p-values extracted under the section wald-type statistics. Conversely, we will interpret p≥.02 as a failure to confirm our hypothesis. In any case, the post hoc power achieved and the Cohen's f effect size will be reported. The effect will be deemed irrelevant if the power achieved is lower than 90% or if the Cohen's f effect size is smaller than the minimum detectable effect size we obtained from the sensitivity analysis.