Defacing biases in manual and automated quality assessments of structural MRI with MRIQC

Céline Provins[1], Elodie Savary[1], Yasser Alemán-Gómez[1,2], Jonas Richiardi[1], Russell A. Poldrack[3], Patric Hagmann[1], Oscar Esteban[1]

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

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

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

Abstract

A critical requirement before data-sharing of human neuroimaging is removing~~prior to data-sharing of human neuroimaging is the removal of~~ 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 may introduce undesired variability into downstream analysis and interpretation. Here, we pre-register a study design to investigate the degree to which the so-called *defacing* alters the quality assessment of T1-weighted images of the human brain from the openly available "IXI dataset" ~~(N=580),~~. The effect of defacing on manual quality assessment will be investigated on a single-site subset of the dataset (N=185). By means of repeated-measures analysis of variance (rm-ANOVA)~~,~~ or linear mixed-effects models in case data do not meet rm-ANOVA's assumptions, we will determine whether four trained human raters' perception of quality is significantly influenced by defacing by comparing their ratings on the same set of images in two conditions: "non-defaced" (i.e.. preserving facial features) and "defaced". ~~(N=185 images per condition).~~ Relatedly, we will also verify that defaced images are systematically assigned higher quality ratings.~~grades on average raters are more optimistic about quality in the defaced set~~. In addition, we will ~~also~~ investigate these biases on automated quality assessments by applying multivariate rm-ANOVA (rm-MANOVA) on the image quality metrics extracted with *MRIQC* on the full IXI dataset (N=580; three acquisition sites). The analysis code, tested on simulated data, is made openly available with this pre-registration report. This study seeks ~~strong~~ evidence of the deleterious effects of defacing on ~~quality assessments of the data~~data quality assessments by humans and machine agents.

Introduction

The removal of facial features —or *defacing*— ~~has become~~ is a ~~necessary step~~ before sharing anatomical images of the brain to protect participants' privacy (Schwarz et al. 2021) in compliance with some local privacy protection regulations, such as the General Data Privacy Regulation (GDPR)[1]

[1] 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) [2016] OJ L 119/1

in Europe or the Health Insurance Portability and Accountability Act (HIPAA)[2] 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 Figure 1). Defacing is ~~therefore~~ a destructive step with the potential to alter the results of downstream processing. ~~For instance, d~~De Sitter et al. (2020) showed that downstream ~~automatic~~ automated analysis methods failed in execution up to 19% of the cases after defacing, as opposed to 2% on non-defaced 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. (2021) likewise showed how ~~T~~these failures propagate and accumulate downstream, leading to substantial changes ~~on~~ in the study outcomes. In a similar approach to our design, Bhalerao et al. (2022) explored the impact of different defacing tools on a subset of image quality metrics (IQMs) automatically generated with *MRIQC* (Esteban et al. 2017). They found that all defacing tools had an impact on a subset of IQMs, and they estimated corresponding effect sizes on a sample limited to 30 subjects with a univariate modeling approach. Moreover, they ~~analyzed~~ identified further effects on the downstream segmentation of images. However, their work did not investigate biases in manual assessment.

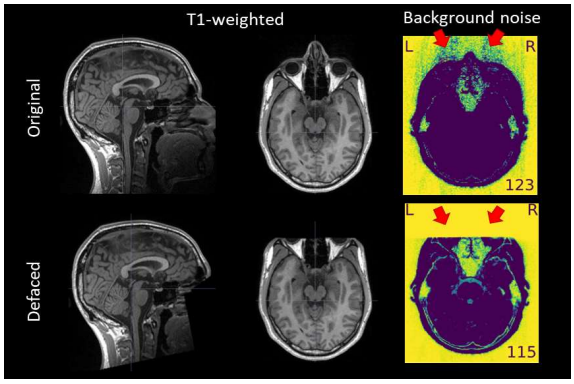

Figure 1. An example of T1w image before and after defacing. Defacing is typically implemented by zeroing the voxels around the face. The background noise visualization is extracted from the *MRIQC* visual report and illustrates that eye spillover is one example of key information in evaluating image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

Formatted: Font: Italic, English (United States)

Here, we set out to understand how defacing influences the outcomes of both manual and automated quality assessment (QA) of unprocessed data (Esteban et al. 2020). This initial QA checkpoint is critical to identify substandard MRI data and exclude them early from the research workflow (which ~~correspond~~ corresponds to performing quality control, QC). Indeed, there is strong evidence that data showing specific artifacts or insufficient overall quality introduce bias into the results of analyses, raising questions about their validity (Power et al. 2012; Zalesky et al. 2016; Alexander-Bloch et al. 2016). As an example, Alexander-Bloch et al. (2016) showed that in-scanner motion can lead to systematic and regionally-specific biases in anatomical estimation of features of interest such as cortical thickness.

The ~~very~~ limited reliability of automated alternatives~~,~~ largely due to site- ~~effects (Esteban, Poldrack, and Gorgolewski 2018),~~ leads to implementing QA manually, by screening the imaging data ~~in~~ on a

[2] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, S. 264

one-by-one basis. However, ~~V~~visual inspection is ~~however~~ time-consuming~~,~~ and prone to large intra- and inter-rater variabilities. Therefore, the implementation of ~~interfaces~~ assisting tools and protocols to efficiently screen and QA large datasets~~overcome such challenges~~, e.g., *MRIQC* (Esteban et al. 2017), *MindControl* (Keshavan et al. 2018), and *Swipes4science* (Keshavan, Yeatman, and Rokem 2019). is an active line of work. Large consortia have also made substantial ~~investment~~ investments in this important task and have generated valuable contributions to QA/QC protocols, e.g., the Human Connectome Project (Marcus et al. 2013), or the INDI initiative (QAP; Shehzad et al. 2015). One related, but conceptually innovative approach was proposed by the UK Biobank (Alfaro-Almagro et al. 2018), where sufficient quality was operationalized as the success of downstream processing. Given the massive size of the UK Biobank, ~~(~~Alfaro-Almagro et al. (2018) flagged for exclusion ~~those~~ the images that did not successfully undergo pre-processing ~~for exclusion~~. Although image exclusions ~~responded~~ 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 (Mortamet et al. 2009; Shehzad et al. 2015; Esteban et al. 2017), or by training deep models directly on 3D images (Garcia, Dosenbach, and Kelly 2022). However, the problem remains extremely challenging when predicting the quality of images acquired at a new center yet unseen by the model (Esteban et al. 2017; Esteban, Poldrack, and Gorgolewski 2018).

In a recent exploration (Provins et al. 2022), we found preliminary evidence that defacing alters both the manual and automatic assessments of T_1-weighted (T1w) MRI images on a small sample (N=10 subjects per defaced/non-defaced condition), implemented with *MRIQC*. The present paper aims ~~at confirming~~to confirm the latter analysis on ~~a~~ larger, unseen, samples (N=185 in the investigation of manual QA; N=580 in automated QA).

Methods

Hypotheses

~~The overarching question behind t~~This pre-registered report ~~is~~sets out to confirm whether defacing alters the manual and automatic ~~assessment~~QA of T1w images of the healthy, human brain, implemented with *MRIQC*. ~~To do so, we will~~This overarching question will be tested in two specific hypotheses~~ ~~:

1. Defacing influences trained experts' perception of quality, ~~and~~leading to significant differences in their ~~their~~quality ratings~~will significantly vary~~ between the defaced and the non-defaced images~~conditions~~. ~~Besides~~Specifically, ~~because there is less information in the image after the removal of facial features,~~ we expect raters ~~will~~to assign ~~more optimistic (better~~higher ~~, on average)~~ ratings, on average, in the defaced condition than in the corresponding non-defaced condition (see Figure 1) ~~; and~~.
2. Defacing influences automatic QA/QC with *MRIQC*, ~~hence it will~~ introduce~~ing a~~ significant and systematic biases in vectors of IQMs ~~computed by *MRIQC* between the~~extracted from defaced and ~~the~~ non-defaced ~~conditions~~images. As evidenced by our preliminary data (Provins et al., 2022), these

Formatted: Font: Italic, English (United States)

Formatted: Font: Italic, English (United States)

Formatted: Subscript

biases may showcase one direction for some IQMs and the opposite or ~~no effects~~no effects on others. Therefore, the directionality of effects cannot be hypothesized.

Data

This confirmatory analysis is based on the publicly available IXI dataset (Hill et al. 2006), which contains 580 non-defaced T1w images acquired at three different sites featuring one 3T (Hammersmith Hospital, London, UK) and two 1.5T devices (Guy's Hospital, London, UK, and Institute of Psychiatry, Psychology & Neuroscience, London, UK). The scanner parameters available for each site are listed in Table S1. None of the authors have screened or queried the dataset to anticipate any quality-related patterns or summary statistics. Moreover, except for author OE, the other authors have neither accessed nor performed any type of processing on the data before pre-registration. One exclusion ~~criteria~~ criterion for the subjects in the IXI dataset will be the absence of a T1w scan. No subjects will be excluded from our analysis on the basis of ~~data~~ quality of the ~~original~~ non-defaced images when evaluating the influence of defacing in automatic QA/QC (hypothesis 3). In the case of hypotheses 1 and 2, experiments will be carried out on the ~~full~~ subset of 185 images acquired at the 3T site (Hammersmith Hospital, London, UK). Images will be excluded from the evaluation of hypotheses 1 and 2 in the case of complete failure of image reconstruction, or if ~~the~~ an image~~s~~ ~~was~~ was assigned the lowest grade (one in our 1-4 interval scale) in both conditions by all raters.

Data processing. First, a defaced version of each scan will be generated with PyDeface (Gulban et al. 2019). ~~PyDeface is chosen~~We chose *PyDeface* because it presents the highest success rate at removing facial features while not removing brain voxels (Theyers et al. 2021). Furthermore, Bhalerao et al. (2022) showed that *PyDeface* resulted in the smallest effect size on the noise-based IQMs.~~nly if PyDeface fails resulting in the preservation of substantial facial features from the original image, will images be excluded from the analysis.~~ No images will be excluded on the grounds of ineffective defacing. Under our hypotheses, images partially retaining facial features (e.g., sections of the eyes and the background around them) are expected to be more consistent between conditions for humans and machines. Therefore, we will not exclude these images despite their potential contribution to reducing effect sizes. ~~To impede the matching~~The raters will assess the quality of the same images in ~~of the~~ two conditions (non-defaced and defaced) ~~for a single individual~~. Raters will not have access to the mapping between defaced and non-defaced counterparts. We will obfuscate participant identifiers and shuffle their ordering before presentation ~~participant identifiers will be randomized under both conditions by reassigning 1240 randomly drawn unique identifiers (580/580 non-defaced/defaced + 40/40 repeated non-defaced/defaced repeated images).~~ *~~MRIQC~~* The latest version in the 2~~.0.6~~3.1 series of *MRIQC* will ~~then~~ be executed on all the T1w images available (that is, non-defaced and defaced). Once all individual image processing with *MRIQC* ~~are~~ is done, the IQMs corresponding to every image in the sample will be collated and converted into a tabular format ~~with *MRIQC*'s "group" processing tool. No images will be excluded on the grounds of ineffective defacing. Under our hypotheses, images partially retaining facial features (e.g., sections of the eyes and the background around them) are expected to be more consistent between conditions, for both~~

Formatted: Font: Italic, English (United States)

Formatted: Font: Italic, English (United States)

~~humans and machines. Therefore, we will not exclude these images despite they might contribute to reducing effect sizes.~~ The local ethics committee has approved the processing of non-defaced images. This study does not attempt to re-identify the participants, nor facilitate in any way such efforts. Should the data of any of the participants be recalled from the original IXI dataset, e.g., after a UK GDPR request, we will accordingly recall the corresponding visual reports generated by *MRIQC*. ~~Processing non-defaced images has been approved by the local ethics committee.~~

Manual assessment protocol. We will perform manual quality assessment only on the images coming from the ~~sole~~ site with a 3 Tesla (3T) device (Hammersmith Hospital; N=185). This choice effectively eliminates the field strength and other variability sources emerging from the specific scanning site as potential random effects. Moreover, images acquired with the 3T scanner are expected to showcase ~~a better~~ signal-to-noise ratio (SNR) twice as high as the SNR of images acquired with 1.5T scanners~~, and~~. Thus, the images acquired with the 3T scanner likely yield, on average, better quality assessments ~~on average~~ by human raters independently of the defacing condition. Four human raters will assess the quality of the subsample, in each of the two conditions (that is, defaced and non-defaced). The quality assessment will be carried out with the individual screening of one *MRIQC*-generated visual report per subject and condition. These reports will be openly shared (see Data and code availability statement). Raters will be recruited by inviting volunteers via e-mail with the mailing list of the Department of Radiology of the Lausanne University Hospital (CHUV, Lausanne, Switzerland). We will not impose restrictions on the experience of the raters beyond familiarity with T1w images of the human brain. To ensure consistency of their training, raters will read our published QC protocol (Provins et al. 2023) and take a 4h training session. At the beginning of this session, the raters will self-assess their experience as either beginner, intermediate or advanced. The materials corresponding to the training session as well as the self-assessments of experience will be openly shared for future exploration (see Data and code availability statement). ~~Furthermore, to~~ To assess the intra-rater effects on QA, 40 ~~images~~ subjects selected randomly will be presented a second time in both conditions to all raters without them knowing it. This sums up to ~~a total of~~ 450 images per rater (225 images per condition). We chose to repeat 40 subjects because it represents a good trade-off between having enough statistical power and the risk of having raters who do not complete their assignment. The random number generator to choose the 40 repeated subjects and the obfuscation of participant identifiers will be initialized with the timestamp of submission and converted to integer with the format YYMMDD + SSmmHH (Y: year, two last digits; M: month, D: day; S: seconds; m: minutes; H: hour). This seed will then be preserved, clearly reported, and set for all the analyses. After screening each visual report, the ~~R~~raters will assign each image a quality grade with the rating widget presented in Figure 2. A quality score will be assigned using a slider that permits the selection of numbers in a continuous scale from 1 to 4 (interval step of 0.05 and 1 corresponding to the lowest quality) ~~(1 : excluded, 4 : excellent quality) with the help of the visual reports generated by *MRIQC*, which was modified to allow~~in order to produce~~ing interval ratings (see Figure 2)~~. As presented in Figure 2, ~~T~~the slider is presented with four categorical ranges ~~categories (1 : excluded, 4 : excellent quality) are shown~~ for reference~~, but the actual rating is not categorical (interval step of 0.05)~~. The starting position of the slider is set in the middle. The raters will be instructed to ~~base their~~

Formatted: Font: Italic, English (United States)

~~quality assessment on~~assess each subject according ~~to the exclusion criteria described in~~ our QC protocol (Provins et al. 2023), and they will not have access to the IQMs. ~~The starting position of the slider is set in the middle.~~ All raters will view the visual reports on a single LED panel of 43" screen diagonal ~~and~~, 3840 × 2160 resolution ~~and~~, 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 takes. The time for each assessment will be measured and made available for future exploration. The~~Raters, the~~ assignment of images in the two conditions to raters, the blinding of image identifiers, the shuffling of presentation, and the tracking of raters' progress will be all managed with a Web Service we have developed for this study called Q'kay. It is described in detail in (Savary et al. 2023).

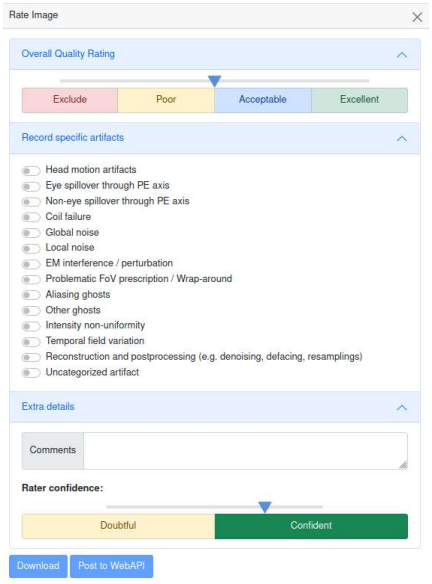

Figure 2. *MRIQC* rating widget has been modified so that quality grades are assigned using a slider. The latter ranges from 1 to 4 ~~(1 : excluded, 4 : excellent)~~ and allows to produce interval ratings. The categories are indicated as hints but the actual rating is fine-grained (interval step of 0.05). Additionally, we added a field to insert comments and a slider to indicate the rater's confidence. The latter is recorded on a scale from 0 to 1 and the categories below the slider are indicated as hints. The rater's confidence and the selected list of artefacts will be recorded and shared ~~in the supplementary material~~ for future exploration, but they will not be accounted for in the confirmatory analyses of this manuscript. These modifications are available in *MRIQC* version 22.0.3 and above.

Formatted: Font: Italic

Experiments

Determining that defacing biases the human raters' assessments on quality. We will test the influence of the defacing condition and the rater (within-subject factor variables) on the ratings (dependent variable) using rm-ANOVA, or linear mixed-effects models in case data do not meet rm-ANOVA's assumptions. As opposed to multiple t-tests, rm-ANOVA and linear mixed-effects models enable ~~to disentangle~~disentangling the variability coming from the raters and the variability coming from defacing and ~~to quantify~~quantifying the ~~latters~~latter. Indeed, because we do not necessarily expect the ratings distribution of each rater to have the same mean, rm-ANOVA and linear mixed-effects models account for the baseline difference in ratings by adding the rater as a random effect in the model. We will first verify that the sphericity and normality assumptions of rm-ANOVA are met. The normality assumption will be verified with the Shapiro-Wilk normality test (Shapiro and Wilk

1965), implemented in the shapiro.test function of the ggpubr R package (Kassambara 2020). Sphericity will be assessed with Mauchly's test for sphericity (Mauchly 1940), implemented in the rstatix R package (Kassambara 2021). Rm-ANOVA will then be implemented with the anova_test function from the rstatix R package and ~~the standard~~a significance level of p<.02 ~~level for significance~~ will be applied. We determined using G*Power (Faul et al. 2009; see Figure 3) that with rm-ANOVA our experimental design can at worst identify effects of f=0.14 corresponding to~~or~~ $\eta^2 = 0.019$ (i.e., a ~~small~~ medium effect) or greater with a power of 90% (see Equation S1 to convert effect size of type f to type η^2). To put this number into perspective, in our pilot study, we found an effect size of f=0.31 (see Equation S2 and S3 for how it was calculated). Comparison between both effect sizes needs however to be performed with caution as the design of the rating collection has been modified between the pilot study and this pre-registration. ~~I~~In the contingency that at least one of the assumptions of rm-ANOVA is violated, ~~rm-ANOVA~~this test will not be employed, and we will use linear mixed-effects models instead. The latter will be~~,~~ implemented in R with the lmer function of the lme4 package (Bates et al. 2022). As part of regression diagnostics, we will examine the shape of the regression residuals, which will be reported in the supplementary materials for completeness ~~to choose an appropriate distribution~~. Indeed, non-Gaussian or~~,~~ heteroscedastic ~~or~~ residuals indicate non-optimal model fit. ~~I~~To test the effect of defacing, we will perform a likelihood-ratio test comparing the linear mixed-effects models with and without adding the defaced factor as a fixed effect. In both compared models, the intercept will be allowed to vary between raters (i.e.~~,~~ the rater factor will be included as a random effect). The likelihood-ratio test will be implemented with the anova function of the R package stats. The bias of defacing on the manual ratings will be deemed significant if the likelihood-ratio test returns p<.02. ~~In addition, we will compute the Bayes Factor between models to obtain a qualitative estimate of the importance of the effect.~~ In addition, to estimate the importance of the effect, we will compute the non-centrality parameter associated with the likelihood ratio test, which is a proxy for its power (Kirk 2012). We will deem the effect irrelevant if the latter parameter is smaller than 13, corresponding to the minimum power achievable from the sensitivity analysis. Lastly, the variance related to the intra-rater effect will be estimated using the rm-ANOVA or, in the contingency case, by computing the variance of the regression coefficients linked to the random effect.

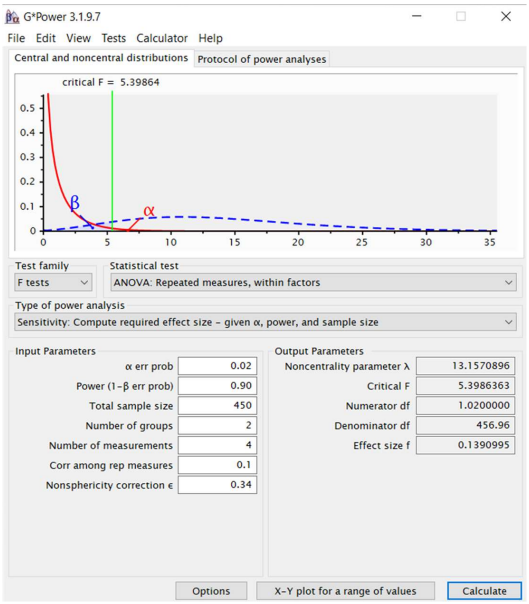

Figure 3. The sensitivity analysis indicates that at worst, using rm-ANOVA, we will be able to confirm differences in manual ratings of f=0.14 corresponding to $\eta^2 = 0.019$ (i.e a medium~~small~~ effect) or greater. We ran a sensitivity analysis with G*Power (Faul et al. 2009) setting ~~. Given that the primary hypothesis has~~ two groups (defaced/non-defaced) and 4 measurements (4 raters) with a total sample size of (185+40)*2 = 450 (number of subjects in Hammersmith Hospital + number of images presented twice, multiplied by the two groups), with 90% power, α = 0.02, a nonsphericity correction of 0.34 and a correlation among repeated measures of 0.1~~, we will be able to confirm differences of f=0.14 (i.e., a small effect) or greater~~. Note that this sensitivity analysis is conservative as we expect the correlation among repeated measures to be much higher, which would reduce the detectable effect size. Furthermore, the lowest sphericity correction possible was used to maximize the detectable effect size.

Formatted Table

Formatted: Font: Bold, English (United States)

226

Confirming that on average ratings are ~~more optimistic~~higher on defaced images. We will use Bland-Altman (BA) plots (Altman and Bland 1983) to visualize the bias and the limits of agreement of manual quality ratings between the non-defaced and the defaced condition. 5 BA plots will be generated and reported either in the supplementary material or the main manuscript: one for each individual rater and one pooling the ratings from all raters together. We will use the BA plots of each individual rater to investigate whether the bias varies with respect to the quality grade attributed and how the bias changes depending on the rater. The BA plot with the pooled ratings will be used to test the significance of the bias. ~~-~~To demonstrate that the ratings of the defaced condition are ~~more optimistic~~higher than the corresponding ratings on the non-defaced 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 (see Figure 4 from our pilot study for reference). If the distribution of ratings is not Gaussian (Shapiro-Wilk test), we will use non-parametric 95% limits of agreement (Bland and Altman 1999). An important difference to note is that the BA plot on Figure 4 has been produced with categorical ratings, unlike the one we plan to generate for this manuscript. The modification of the ratings from categorical to interval stems from the impossibility of running proper statistical tests on the rating design of our pilot study.~~Furthermore, we will investigate whether the bias varies with respect to the quality grade attributed and how the bias changes depending on the rater.~~

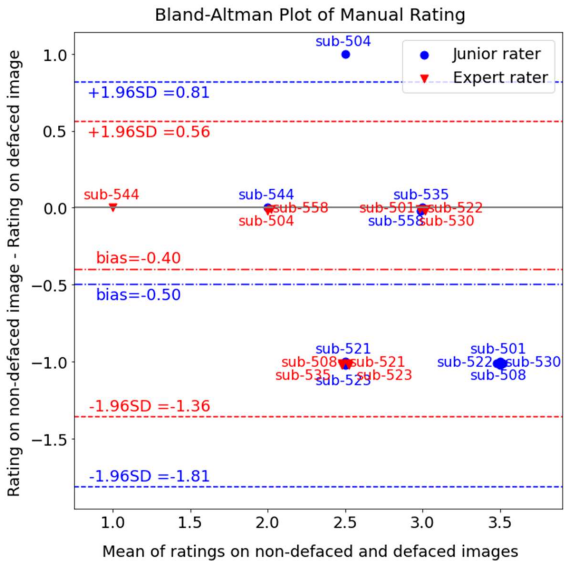

Figure 4. The BA plot from our pilot study showcasing manual ratings. Note that this plot from our pilot study, unlike the one we plan to generate for this manuscript, has been produced with categorical ratings, hence its discrete appearance. The bias is determined by computing the mean of differences and visualized by placing a dashdotted line at that value. The 95% confidence interval is constructed as the bias ± 1.96 * the standard deviation of the differences. It is represented by two dashed lines. To highlight the situation where non-defaced/defaced images were assigned the same quality score, a full line is placed at the zero difference. Ratings are annotated with the corresponding subject identifier to allow further exploration.

Formatted Table

Determining that defacing introduces biases in *MRIQC*-generated IQMs. Defacing impact on automatic QA will be evaluated based on the 62 IQMs calculated by *MRIQC*. For the complete list of IQMs produced by *MRIQC* and their definitions, refer to Table 2 in (Esteban et al. 2017). A two-way repeated-measures MANOVA (rm-MANOVA) will be used to test whether defacing significantly influences the IQMs. This test will be implemented with the multRM function of the MANOVA.RM package in R. However, because many IQMs are heavily correlated (see Figure 5), reducing the dimensionality of the IQMs before applying rm-MANOVA is necessary. We will thus apply principal components analysis (PCA) on the IQMs. Specifically, PCA will be applied only on the IQMs coming from the ~~original~~ non-defaced data and the resulting transformation will be applied to IQMs coming both from the ~~original~~ non-defaced and defaced data. Performing PCA only on the IQMs coming from the ~~original~~ non-defaced data is essential to ensure the defacing effects are not mitigated. PCA will be implemented with the prcomp function of the stats package of R, with the option scale=TRUE meaning that the variables are standardized to have unit variance before the decomposition. The number of principal components will be determined by the Kaiser criterion, and thereby we will keep components with an eigenvalue above 1.0. Consequently, the rm-MANOVA will be constructed with the projected IQMs as the continuous dependent variables and two categorical independent variables, one corresponding to the (non-defaced or defaced) condition of the image, the other corresponding to

the scanning site. Adding the scanning site as an independent variable allows us to control for differences in IQMs that arise from site-effects (Esteban, Poldrack, and Gorgolewski 2018, Morgan et al. 2022). We will apply ~~the standard~~a significance level of p<.02 ~~level for significance of~~ for the rm-MANOVA and consider the p-values extracted under the Wald-type statistics section. We determined using G*Power (Faul et al. 2009; see Figure 6) that our experimental design can identify, with a 90% power, effects of f=0.16 corresponding to~~or~~ $\eta^2 = 0.025$ (i.e., a medium~~small~~ effect) or greater. To put this number into context, the effect size associated with the MANOVA on the IQMS of our pilot study was f=0.16 (see Equation S4 and S5 for its computation). Comparison between both effect sizes needs however to be exercised with caution as the statistical design are different; in our pilot study, we used a normal MANOVA on only 5 IQMs that showed the strongest bias on the BA plot while in this pre-registration we are planning to use a repeated-measures MANOVA with all IQMs projected onto the PCA basis. ~~For reference~~In addition, the effect size associated with PyDeface influence on IQMs in (Bhalerao et al. 2022) ranged from ~~0.09 to 3.58~~f=0.045 to f=1.79 with a mean effect size across IQMs of ~~1.23~~f=0.61 (see Eequation S3 for the conversion of Cohen's d to Cohen's f). ~~Furthermore~~To visualize the defacing bias on the automatic quality ratings, we will ~~also also visualize the IQM with BA plots (as described above). A grid of 62 BA plots, one per IQMs, will be generated~~generate a BA plot (as described above) for each IQM and for each principal component. All BA plots will be reported in the supplementary material, and the ones that are most clear, interpretable and descriptive will be presented in the main manuscript.

Data and code availability statement

The IXI dataset is available at https://brain-development.org/ixi-dataset/ ~~(URL)~~ under the Creative Commons CC BY-SA 3.0 license. The IQMs that we used to create Figure 5 were extracted from all the available T1w images of the ABIDE dataset, and are openly available within the *MRIQC*-learn package. The Web Service that we implemented to collect the manual ratings in this study is available under the Apache 2.0 license at https://github.com/nipreps/qkay. All the new materials relating to this work will be 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. Material that could qualify as adaptations of the original IXI dataset (that is, the individual reports generated by *MRIQC*) will be released under the terms of the CC-BY-SA-4.0 license.

~~Before publication, we have initiated a "CodeOcean Capsule" to provide reviewers with private and anonymous access to the source code for peer-review, which can be accessed at https://codeocean.com/capsule/8731863/tree.~~

Conclusion

This study is proposed to investigate whether manual and automatic aspects of QA/QC implemented in *MRIQC* are biased by the process of defacing data. We plan to openly share all the materials under suitable licenses upon publication. ~~(Apache 2.0 for code and CC-BY for data) upon publication. Before publication, we have initiated a "CodeOcean Capsule" to provide reviewers with private and anonymous access to the source code for peer-review, which can be accessed at~~

Formatted: Space After: 12 pt

Formatted: Font: Italic, English (United States)

https://codeocean.com/capsule/8731863/treesee Data and code availability statement).

FinallyMoreover, a discussion has been included within the supplementary material, speculating the impact of this study should the hypotheses be verified.

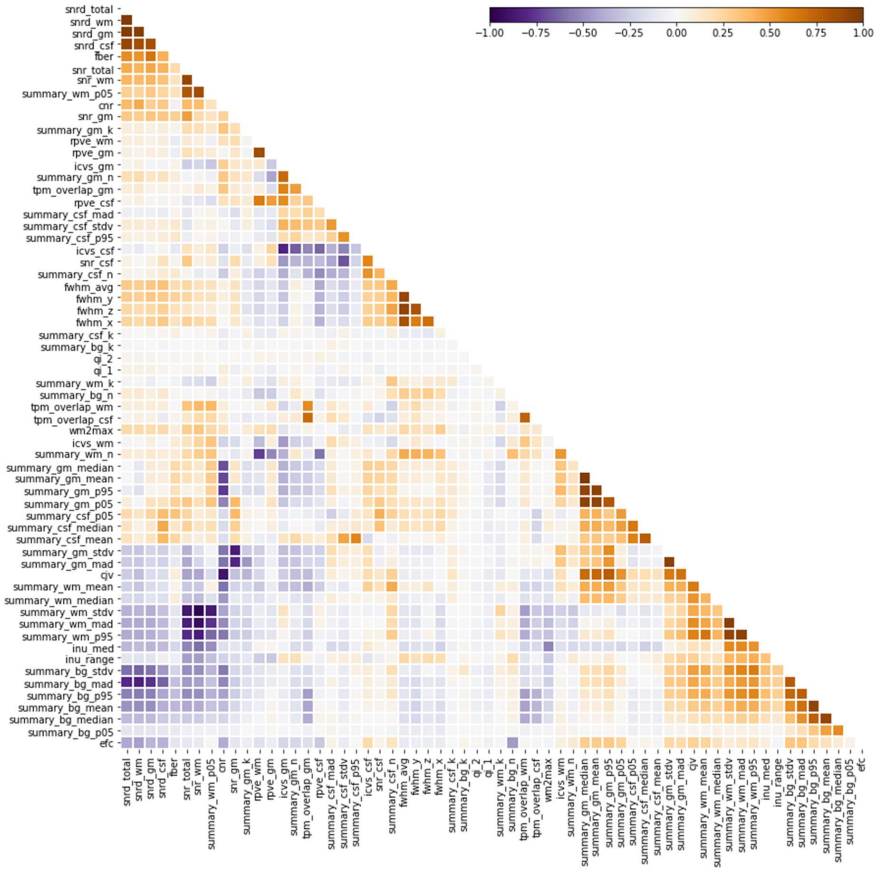

Figure 5. A number of IQMS are highly correlated. IQMs were extracted from all the available T1w images of the ABIDE dataset, and are openly available within the *MRIQC*-learn package. We performed hierarchical clustering on the correlation plot to visualize more clearly clusters of correlated IQMs. A similar plot will be generated for this study.

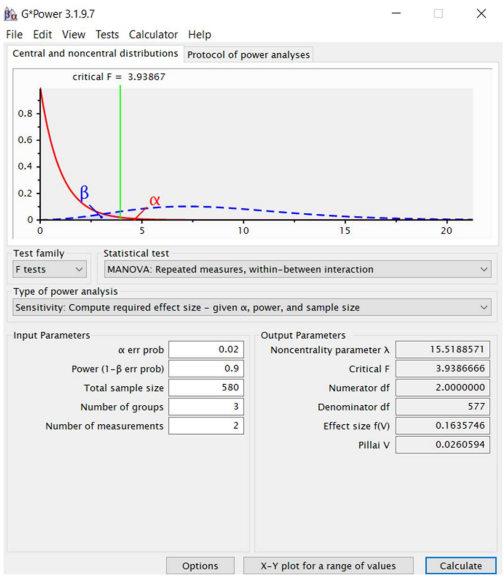

Figure 6. The sensitivity analysis indicates that we will be able to confirm differences in IQM of f=0.16 corresponding to $\eta^2 = 0.025$ (i.e a ~~small~~ medium effect) or greater. We ran a sensitivity analysis with G*Power (Faul et al. 2009~~). Given that the primary hypothesis has~~) setting three groups (3 sites) and 2 measurements (defaced/non-defaced) with N = 580 (number of T1w per subject) per condition, with 90% power, and α = 0.02~~, we will be able to confirm differences of f=0.16 (i.e., a small effect) or greater~~.

Table 1. 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 ~~experts'~~ raters' perception of quality	Do the quality ratings from human raters significantly vary between the defaced and the non-defaced 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 Figure 3, 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 ~~small~~ 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 non-defaced conditions. 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.
			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 Figure S1.	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≥.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 ~~more optimistic (better, on average)~~ than the correspondi ng ratings on the non-defaced 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 non-defaced and the defaced condition.		To demonstrate that the ratings of the defaced condition are ~~more optimistic~~higher than the corresponding ratings on the non-defaced 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 non-defaced 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 ~~0.09 to 3.58~~ with a mean effect size across IQMs of ~~1.23~~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 5), we will apply principal components analysis (PCA) on the IQMs before running rm-MANOVA.	The sensitivity analysis, reported in Figure 6, indicates that we will be able to confirm differences in IQM of f=0.16 corresponding to $\eta^2 = 0.025$ (i.e a ~~small~~ 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 non-defaced 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.

Acknowledgments

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

References

Alexander-Bloch, Aaron, Liv Clasen, Michael Stockman, Lisa Ronan, Francois Lalonde, Jay Giedd, and Armin Raznahan. 2016. “Subtle In-Scanner Motion Biases Automated Measurement of Brain Anatomy from in Vivo MRI.” *Human Brain Mapping* 37 (7): 2385–97. https://doi.org/10.1002/hbm.23180.

Alfaro-Almagro, Fidel, Mark Jenkinson, Neal K. Bangerter, Jesper L.R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N. Sotiropoulos, et al. 2018. “Image Processing and Quality Control for the First 10,000 Brain Imaging Datasets from UK Biobank.” *Neuroimage* 166 (February): 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034.

Altman, D. G., and J. M. Bland. 1983. “Measurement in Medicine: The Analysis of Method Comparison Studies.” *Journal of the Royal Statistical Society. Series D (The Statistician)* 32 (3): 307–17. https://doi.org/10.2307/2987937.

Bates, Douglas, Martin Maechler, Ben Bolker [aut, cre, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, et al. 2022. “Lme4: Linear Mixed-Effects Models Using ‘Eigen’ and S4.” https://CRAN.R-project.org/package=lme4.

Bhalerao, Gaurav Vivek, Pravesh Parekh, Jitender Saini, Ganesan Venkatasubramanian, John P. John, Biju Viswanath, Naren P. Rao, et al. 2022. “Systematic Evaluation of the Impact of Defacing on Quality and Volumetric Assessments on T1-Weighted MR-Images.” *Journal of Neuroradiology* 49 (3): 250–57. https://doi.org/10.1016/j.neurad.2021.03.001.

Bland, J Martin, and Douglas G Altman. “Measuring Agreement in Method Comparison Studies.” Statistical Methods in Medical Research 8, no. 2 (April 1, 1999): 135–60. https://doi.org/10.1177/096228029900800204.

Esteban, Oscar, Daniel Birman, Marie Schaer, Oluwasanmi O. Koyejo, Russell A. Poldrack, and Krzysztof J. Gorgolewski. 2017. “MRIQC: Advancing the Automatic Prediction of Image Quality in MRI from Unseen Sites.” Edited by Boris C Bernhardt. *PLOS ONE* 12 (9): e0184661–e0184661. https://doi.org/10.1371/journal.pone.0184661.

Esteban, Oscar, Rastko Ciric, Karolina Finc, Ross W. Blair, Christopher J. Markiewicz, Craig A. Moodie, James D. Kent, et al. 2020. “Analysis of Task-Based Functional MRI Data Preprocessed with FMRIPrep.” *Nature Protocols* 15 (7): 2186–2202. https://doi.org/10.1038/s41596-020-0327-3.

Esteban, Oscar, Russell A. Poldrack, and Krzysztof J. Gorgolewski. 2018. “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*, edited by Danail Stoyanov, Zeike Taylor, Simone Balocco, Raphael Sznitman, Anne Martel, Lena Maier-Hein, Luc Duong, et al., 190–99. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-01364-6_21.

Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. “Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses.” *Behavior Research Methods* 41 (4): 1149–60. https://doi.org/10.3758/BRM.41.4.1149.

Garcia, Mélanie, Nico Dosenbach, and Clare Kelly. 2022. “BrainQCNet: A Deep Learning Attention-Based Model for Multi-Scale Detection of Artifacts in Brain Structural MRI Scans.” bioRxiv. https://doi.org/10.1101/2022.03.11.483983.

Gulban, Omer Faruk, Dylan Nielson, Russ Poldrack, John Lee, Chris Gorgolewski, Vanessa Sochat, and Satrajit Ghosh. 2019. “Poldracklab/Pydeface: V2.0.0.” Zenodo. https://doi.org/10.5281/zenodo.3524401.

Hill, D., S. Williams, D. Hawkes, and S. Smith. 2006. “IXI Dataset – Brain Development.” 2006. https://brain-development.org/ixi-dataset/.

Kassambara, Alboukadel. 2020. “Ggpubr: ‘ggplot2’ Based Publication Ready Plots.” https://CRAN.R-project.org/package=ggpubr.

———. 2021. “Rstatix: Pipe-Friendly Framework for Basic Statistical Tests.” https://CRAN.R-project.org/package=rstatix.

Keshavan, Anisha, Esha Datta, Ian M. McDonough, Christopher R. Madan, Kesshi Jordan, and Roland G. Henry. 2018. “Mindcontrol: A Web Application for Brain Segmentation Quality Control.” *NeuroImage*, Segmenting the Brain, 170 (April): 365–72. https://doi.org/10.1016/j.neuroimage.2017.03.055.

Keshavan, Anisha, Jason D. Yeatman, and Ariel Rokem. 2019. “Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging.” *Frontiers in Neuroinformatics* 13. https://www.frontiersin.org/article/10.3389/fninf.2019.00029.

Kirk, Roger E. 2012. “Experimental Design: Procedures for the Behavioral Sciences. ” *SAGE Publications*.

Marcus, Daniel S., Michael P. Harms, Abraham Z. Snyder, Mark Jenkinson, J. Anthony Wilson, Matthew F. Glasser, Deanna M. Barch, et al. 2013. “Human Connectome Project Informatics: Quality Control, Database Services, and Data Visualization.” *NeuroImage*, Mapping the Connectome, 80 (October): 202–19. https://doi.org/10.1016/j.neuroimage.2013.05.077.

Mauchly, John W. 1940. “Significance Test for Sphericity of a Normal n-Variate Distribution.” *The Annals of Mathematical Statistics* 11 (2): 204–9. https://doi.org/10.1214/aoms/1177731915.

Morgan, Catherine A., Reece P. Roberts, Tessa Chaffey, Lenore Tahara-Eckl, Meghan van der Meer, Matthias Günther, Timothy J. Anderson, et al. “Reproducibility and Repeatability of Magnetic Resonance Imaging in Dementia.” Physica Medica 101 (September 1, 2022): 8–17. https://doi.org/10.1016/j.ejmp.2022.06.012.

Mortamet, Bénédicte, Matt A. Bernstein, Clifford R. Jack, Jeffrey L. Gunter, Chadwick Ward, Paula J. Britson, Reto Meuli, Jean-Philippe Thiran, Gunnar Krueger, and Alzheimer’s Disease Neuroimaging Initiative. 2009. “Automatic Quality Assessment in Structural Brain Magnetic Resonance Imaging.” *Magnetic Resonance in Medicine* 62 (2): 365–72. https://doi.org/10.1002/mrm.21992.

Power, Jonathan D., Kelly A. Barnes, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. 2012. “Spurious but Systematic Correlations in Functional Connectivity MRI Networks Arise from Subject Motion.” *NeuroImage* 59 (3): 2142–54. https://doi.org/10.1016/j.neuroimage.2011.10.018.

Provins, Céline, Yasser Alemán-Gómez, Martine Cleusix, Raoul Jenni, Jonas Richiardi, Patric Hagmann, and Oscar Esteban. 2022. “Defacing Biases Manual and Automated Quality Assessments of Structural MRI with MRIQC.” *28th Annual Meeting of the Organization for Human Brain Mapping (OHBM)* Glasgow, UK. p. WTh566. https://doi.org/10.31219/osf.io/8mcyz.~~Provins, Céline, Yasser Alemán-Gómez, Martine Cleusix, Raoul Jenni, Jonas Richiardi, Patric Hagmann, and Oscar Esteban. 2022. “Defacing Biases Manual and Automated Quality Assessments of Structural MRI with MRIQC.” OSF Preprints. .~~

Provins, Céline, Eilidh MacNicol, Saren H. Seeley, Patric Hagmann, and Oscar Esteban. 2023 “Quality Control in Functional MRI Studies with MRIQC and FMRIPrep.” *Frontiers in Neuroimaging* 1. https://doi.org/10.3389/fnimg.2022.1073734.

Savary, Elodie, Céline Provins, Thomas Sanchez, and Oscar Esteban. 2023. “Q’kay: A Manager for the Quality Assessment of Large Neuroimaging Studies.” *29th Annual Meeting of the Organization for Human Brain Mapping (OHBM). Montreal, Canada.* https://doi.org/10.31219/osf.io/edx6t.

Schwarz, Christopher G., Walter K. Kremers, Heather J. Wiste, Jeffrey L. Gunter, Prashanthi Vemuri, Anthony J. Spychalla, Kejal Kantarci, et al. 2021. "Changing the Face of Neuroimaging Research: Comparing a New MRI de-Facing Technique with Popular Alternatives." *NeuroImage* 231 (May): 117845. https://doi.org/10.1016/j.neuroimage.2021.117845.

Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)†." *Biometrika* 52 (3–4): 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Shehzad, Zarrar, Giavasis Steven, Li Qingyang, Benhajali Yassine, Yan Chaogan, Yang Zhen, Milham Michael, Bellec Pierre, and Craddock Cameron. 2015. "The Preprocessed Connectomes Project Quality Assessment Protocol - a Resource for Measuring the Quality of MRI Data." *Frontiers in Neuroscience* 9. https://doi.org/10.3389/conf.fnins.2015.91.00047.

Sitter, A. de, M. Visser, I. Brouwer, K. S. Cover, R. A. van Schijndel, R. S. Eijgelaar, D. M. J. Müller, et al. 2020. "Facing Privacy in Neuroimaging: Removing Facial Features Degrades Performance of Image Analysis Methods." *European Radiology* 30 (2): 1062–74. https://doi.org/10.1007/s00330-019-06459-3.

Theyers, Athena E., Mojdeh Zamyadi, Mark O'Reilly, Robert Bartha, Sean Symons, Glenda M. MacQueen, Stefanie Hassel, et al. 2021. "Multisite Comparison of MRI Defacing Software Across Multiple Cohorts." *Frontiers in Psychiatry* 12: 189. https://doi.org/10.3389/fpsyt.2021.617997.

Zalesky, Andrew, Alex Fornito, Luca Cocchi, Leonardo L. Gollo, Martijn P. van den Heuvel, and Michael Breakspear. 2016. "Connectome Sensitivity or Specificity: Which Is More Important?" *NeuroImage* 142 (November): 407–20. https://doi.org/10.1016/j.neuroimage.2016.06.035.