Authors' Response to Reviews of

C. Provins et al. – "Defacing biases in manual and automated quality assessments of structural MRI with MRIQC"

AR: Authors' Response

Please note that the line numbers refer to the new document without changes highlighted.

1. Reviewer #1

R#1.1:

This preregistered analysis plan describes the approach for evaluating any bias in structural MRI image quality from defacing. The planned study is timely and will have considerable impact given the current focus on data sharing, open science and in turn push-backs to this concerning participant confidentiality. A real strength of the study is that both manual (person) and automated (computer) image quality assessment methods will be assessed, and in a reasonably large data set from multiple sites. The analysis plan is sound and comprehensive. The statistical methods are well described and justified. However I think it could benefit though from some clarity on the methodology and minor issues as listed below.

AR: We thank the reviewer for their valuable feedback and enthusiastic support.

1.1. Methodological issues

R#1.2:

It's not clear why the position is taken that defaced images will be perceived as having higher quality? In the hypotheses section it is stated that "less information in the image after the removal of facial features, raters will assign more optimistic (better, on average) ratings" Why? Do you mean due to the absence of noisy pixels (or ghosting or signal drop off at the surface)? if so I would make this clearer and refer to Fig 1 more. Also the use of "optimistic" about the raters opinion on quality of the data used throughout is a bit vague/subjective – I think better to refer to it simply as a higher quality score.

AR: We appreciate this comment and have edited the manuscript accordingly. We have made a clearer statement in the caption of Figure 1 that we hypothesize defaced images will receive higher (rather than "more optimistic") quality ratings on average because the noise in the background surrounding the face is key when determining the image's quality, however absent in defaced images.

Corresponding edits: Page 1, ll.22; Page 2, caption of Figure 1; Page 3, ll.87; Page 8, ll. 202; Page 14, table 1.

1.2. For the manual ratings

R#1.3:

The inclusion of image repeats to assess intra-ratter reliability is a good addition to the study design, but why 40? please provide more detail (to enable replication)

AR: We acknowledge the importance of justifying the sample size and have accordingly added the explanation to the paper.

Corresponding edits: Page 5, ll.140-142

R#1.4:

How are raters instructed on how to score – 1 is exclude, 4 is excellent quality but are they given any other instructions and/or criteria? What if one rater looks at SNR and the other favours distortion? Presumably there are some standard aspects you will ask them to consider?

AR: We thank both reviewers for raising this important issue. We have added details in the Manual assessment protocol section regarding the criteria we will instruct the raters to follow.

Corresponding edits: Page 5, ll. 156-157

R#1.5:

How many categories are there? The 1 to 4 scale suggests only 4 possible ratings, but the fact a sliding scale bar is provided in the widget and that Figure 4 shows that the junior rater gives sub-504 a 2.5 score suggests non-inter ratings are possible. Pls clarify.

AR: We agree with the reviewer that this point deserved attention, and the text has been modified accordingly. We have also clarified the interpretation of Figure 4.

Corresponding edits: Page 5, ll.156-157; Page 6, caption of Figure 2; Page 8, ll.214-217; Page 8, caption of Figure 4.

R#1.6:

What is the experience of the raters? If they are not known yet, will there be any requirements for the raters previous experience?

AR: We acknowledge the relevance of the previous rater experience, and we have described how we are going to recruit and train raters, as well as how we plan to record their experience.

Corresponding edits: Page 4-5, ll.138-143.

R#1.7:

What are the "Visual reports" the raters will see? P4 L123, L124 To me "reports" suggests results - will they see the automated quality results before their rating? This would likely influence their decision. Or do you simply mean the scoring widget presented to them (fig 2)? But L128 suggests describing how the "reports" are presented suggestd the report is the image iteself ? This needs to

be clearer what the raters are presented with at the time of scoring. The same ambient lighting etc would also be useful to state (as well as display screen)

AR: We have reordered and rewrote the "Manual assessment protocol" section to make it clearer and complete it with the missing information.

Corresponding edits: Page 4, l.135-136; Page 5, ll.151-152; Page 5, ll. 159; Page 9, ll. 259-261.

R#1.8:

How are the confidence measures to be analysed/included? Is it binary or scale bar in Fig 2 implies a range? Are you expecting an interaction with the numeric quality score?

AR: We thank the reviewer for bringing up this interesting question, which is intimately related to R#2.5. Indeed, we are not planning any confirmatory analyses using the rater confidence. However, we will make the data publicly available, and understand that rater confidence may potentially be included as a moderator if rater experience has also large effects. This discussion has been introduced in the supplementary materials, where we have included some speculative discussion of the study.

Corresponding edits: Page 6, caption Figure 2; Supplementary material, page 4.

R#1.9:

How are the biases for raters (from the Bland-Altmann plots) to be used – combined in some way or just presented for each rater?

AR: We thank the reviewer for this relevant question and have accordingly added details to the manuscript.

Corresponding edits: Page 8, ll. 204-209.

1.3. For the automated IQM analysis

R#1.10:

Number of sites: I would consider splitting the currently proposed 3 site single analysis in to separate analyses - since the two 1.5T sites are likely to have different quality metrics from the 3T site, any site effects including 3 site will be a mix of field strength and other site variability. I would add some detail on the approximate scan time, voxel size and head coils used at each site as these will all impact the SNR. The scanner manufacturer and model and sequence used at each site would also be of interest (since you would expect less variability with matched models and sequences). (If of interest, we recently compared MRI measures at different field strength and sites 10.12688/f1000research.20496.1 and 10.1016/j.ejmp.2022.06.012 respectively).

AR: We thank the reviewer for bringing up this important consideration and for sharing these interesting publications. As already described in the paper, we will account for scanning site differences by adding the site as an independent variable in the repeated-measures MANOVA. Nonetheless, we added a sentence to ensure this point is clear in the manuscript and we cited 10.1016/j.ejmp.2022.06.012 as support to the existence of site-effects influencing IQMs. Further-

more, we added all the scanning information available in the supplementary material. Note that the type of sequence, the head coils used and other details are unfortunately not available.

Corresponding edits: Page 3, ll.97-98; Page 9, ll. 232-236, Supplementary Material, Page 1.

R#1.11:

Please specify the total number of IQMS before PCA – 62? Related to this point, above, it's stated that 62 BA plots for the IQM will be generated. How will these results be combined/summarised? Would it be more succinct to only generate BA and report plots for the IQMS identified by PCA?

AR: We thank the reviewer for this suggestion and have incorporated it into the manuscript along with the specification of the total number of IQMs.

Corresponding edits: Page 9, ll. 219; Page 9, ll. 248-251

R#1.12:

What does "As part of regression diagnostics, we will examine the shape of regression residuals to choose an appropriate distribution" please expand/clarify P5 L151 mean?

AR: The reviewer is absolutely right about this problem, as our statistical modeling plan would not allow us to *choose an appropriate distribution*. We have edited the manuscript to clearly indicate that we will comprehensively report the results, including the regression residuals, because their shape can be interpreted as a measure of "model misspecification".

Corresponding edits: Page 7, ll. 188-190.

R#1.13:

Computing the Bayes Factor of the models P6 L158 should allow you to say something quantitative rather than qualitative about the results?

AR: We agree with the reviewer's view, which has led us to completely reconsider the commitment to computing Bayes Factors between models. Instead, we will calculate the non-centrality parameter associated with the likelihood ratio test to quantify the models' fitness.

Corresponding edits: Page 7, ll.195-199; Page 12, table 1.

1.4. Other

R#1.14:

Is significance levels of $p<0.02$ standard? I've not seen it used as the default threshold before. Fine to use this, but maybe remove "standard" on p5 L147

AR: We thank the reviewer for spotting this inconsistency and have rephrased the sentences containing the expression "the standard $p<.02$ significance level".

Corresponding edits: Page 7, ll.195; Page 9, ll.237.

R#1.15:

Is ethical approval in place for data access and analysis (especially considering access to non-defaced data)

AR: We appreciate the comment, and we have more explicitly stated the ethical approval for data access and analysis.

Corresponding edits: Page 4, ll.121-125.

R#1.16:

*Stated in the abstract the study will provide "strong" evidence for or against the deleterious effects of defacing – but the power calculations provided (was useful to see the G*Power outputs, thank you) Cohen's F of > 0.14 may be detected (a small effect or more). Please remove "strong" from the abstract wording.*

AR: We thank the reviewer for spotting this misleading statement and have accordingly removed the "strong" qualifier from the abstract.

Corresponding edits: Page 1, ll.26-27.

1.5. Minor Issues

R#1.17:

Please specify what type of automatic analysis methods/steps Sitter et al. found failed with defaced images (P1, L33).

AR: We now provide further detail about Sitter's report.

Corresponding edits: Page 2, ll. 35-39.

R#1.18:

Qualify what the "very limited reliability" of automated methods is being referred to P2, L50

AR: In agreement with this comment, we have made a more specific description of why the reliability of automated approaches has proved "very limited".

Corresponding edits: Page 2, ll. 55.

R#1.19:

The two conditions: non-defaced and defaced are later referred to as "original" and de-faced. Keeping the same terminology throughout is better. Non-defaced is less ambiguous.

AR: We agree with the reviewer on the importance of keeping a consistent terminology and have replaced all instances of "original" with "non-defaced".

Corresponding edits: Page 4, ll.102; Page 8, ll.226-227; Page 8, axes label of Figure 4.

R#1.20:

The text needs to be checked for grammatical errors, clearer meaning, e.g.,

- *P2 starting on L59, "Given the. . .*
- *P2 starting L61 - "responded" she be responded might be better as "related"*
- *P5 L138*

AR: We thank the reviewer for suggesting text improvements. We have corrected the sentences.

Corresponding edits: Page 3, ll.65-66; Page 6, ll. 169-171.

R#1.21:

The hypotheses section could be clearer –

- *P3 L76 the phrasing "To do so.." does not follow on from the overarching research "question".*
- *P3 L79 – "Besides" would be better replaced with "specifically" since this sentence is now specifically making a hypothesis on the direction of the variance between defaced and non-defaced*

AR: We thank the reviewer for these suggestions and have adapted the text accordingly.

Corresponding edits: Page 3, ll. 83-84; Page 3, ll. 86-87.

R#1.22:

Institute of Psychiatry is now the Institute of Psychiatry, Psychology & Neuroscience

AR: We thank the reviewer for this remark and have corrected the text accordingly.

Corresponding edits: Page 3, ll. 97.

R#1.23:

Some text in the Data processing section P3 L98 and manual assessment section P4 L109 could be reordered for clarity – e.g. the 40 repeat images and two conditions are more clearly introduced in the latter section

AR: We have rewritten the Data processing section for more clarity.

Corresponding edits: Page 4, ll.108-125.

R#1.24:

P4 L112 – SNR at 3T is theoretically double that of 1.5T, so you could simply state that rather

AR: We thank the reviewer for this suggestion and have stated this fact explicitly.

Corresponding edits: Page 4, ll. 129-131.

R#1.25:

P4 L127 – effortless might be better word than "costless"

AR: We thank the reviewer for this suggestion. The text has been corrected.

Corresponding edits: Page 5, ll. 161-164.

R#1.26:

Figure 5

- *The scale bar looks like the max is 100, but it must be 1.00 for the correlation value? Maybe increase the legend text size to make it more readable.*
- *the order of the components can be sorted not alphabetically like shown, but with hierarchical clustering - this would visualise the correlated IQMSs better*
- *make it clearer this is an example of correlations from a different study - ABIDE (What N?) but the same type of plot will be generated for this study*

AR: We thank the reviewer for these interesting suggestions. We have implemented them and replaced Figure 5 with the IQMs correlation plot clustered.

Corresponding edits: Page 11, caption of Figure 5.

2. Reviewer #2

R#2.1:

This registered report describes an experiment to assess the impact of defacing on manual and automated assessments of MR (brain) image quality. A lack of information about image quality can impact downstream processing and introduce bias to further analysis. Understanding the impact of defacing is important as this practice is essential to maintain the privacy of our participants while maximising reuse potential.
The registered report follows the publication of a pilot investigation (Provins, 2022) and extends the sample size from n=10 to n=185. Hypotheses are clear for both measures, but there is no directionality or Bayesian tests described for the analysis of MRIQC IQMs. The publication of the pilot study, code, reuse of an open dataset and submission of a registered report demonstrates exemplary use of this scientific process. The authors should be commended for their commitment. This is an important and interesting question which is described well. My comments relate in places to the positioning and reporting of the study itself, rather than the research methodology, which I think is sound. Below I have commented on the submission in the order which the material was presented. I have reviewed the main submission and supplementary materials, but I was not able to review the code.

AR: We thank the reviewer for the valuable feedback and their enthusiastic support. We also commend their commitment to review not only the manuscript but also the supplementary material and the code.

R#2.2:

Sample size: The analysis is to be conducted on a subsample of all images in the IXI Dataset (n = 185 / 580). References to "n = 580" in the abstract and elsewhere should be removed as they are misleading.

AR: We agree with the reviewer that referencing "N=580" was confusing. We have revised the abstract to unambiguously reflect the sample size of each of the experiments (185 when investigating the manual assessments, 580 when analyzing the image quality metrics extracted automatically).

Corresponding edits: Page 1, ll.14-25.

R#2.3:

Defacing has become necessary "in compliance with privacy regulations" (line 30-31). Not in all regions. Perhaps caveat with "local" and "in some areas", or better still list a few regulations, e.g. GDPR and equivalents. This is an important inclusion to give context to the motivation to understand these effects.

AR: We thank the reviewer for correcting this statement and have adapted it accordingly.

Corresponding edits: Page 1, ll.31-32.

R#2.4:

*It would be useful to highlight more explicitly how the raters (and MRIQC) are *interpreting* the removal of data by defacing. The removal of the eyes as a source of artifact is proposed as one mechanism. Are there any other signals if image quality used or altered? It would be helpful to include a more explicit discussion of this and what 'removal of information' actually means in this context.*

AR: In agreement with Comment R#1.2, we have made a clearer description of how we expect the ratings of defaced images to be higher than those of nondefaced counterparts. To address the second point about the meaning of "removal of information" by defacing, the supplementary materials have been edited to include such a discussion.

Corresponding edits: Supplementary material, page 2.

R#2.5:

The MRIQC rating widget contains a list of artefacts which raters can identify. How was this list developed/tested? How will the data acquired here be tested? State that there will be no analysis of these data or only exploratory analysis of this.

AR: We resonate with the relevance of this assessment, and now we explicitly state that the selected list of artifacts will not be included in our confirmatory analysis. This aspect relates to the later comment 2.14 from this reviewer, and 1.8 from R#1, posing similar questions regarding the collection of rater confidence. As for the other reviewer's comment, we will openly share the data including all these metadata, and report any exploratory analyses that may provide a better insight into the results of the confirmatory analyses.

Corresponding edits: Page 6, caption of Figure 2.

R#2.6:

Hypothesis 2: "Defacing [will] introduce bias in vectors of IQMs computed by MRIQC between the defaced and the non-defaced conditions." (line 82-83). Do you have hypotheses regarding directionality or any IQMs?

AR: We have more clearly stated that we are not hypothesizing about the directionality of the effects on any of the IQMs, and included a justification for such a decision. We hope this study offers sufficient evidence to support more targetted hypothesis about any IQMs or decompositions (i.e., PCA) thereof.

Corresponding edits: Page 3, ll.90-92; Supplementary material, page 2.

R#2.7:

Data: Please state the licence or data usage agreement that the IXI Dataset is made available under.

AR: We have added a comprehensive data availability statement resolving this oversight, and also addressing the related Comment R#2.17.

Corresponding edits: Page 9-10, ll. 252-261.

R#2.8:

Manual assessment procedure - Really good description of the manual assessment protocol, and it appears to be a robust procedure.

AR: We thank the reviewer for their enthusiastic support. To reflect recent advances in the implementation of the rating management tool, we have updated the related text with an updated description of the protocol.

Corresponding edits: Page 4, ll. 116-118; Page 5, ll. 161-164.

R#2.9:

Would be useful to have an estimate of how long each rater will be working for, to guage the workload involved in a replication.

AR: We agree this is a very relevant factor to report, and note that in Page 5, l. 159-160, the manuscript already stated that we will "record the exact time each assessment took". We will also openly share these metadata (as stated in previous comments) and report summary statistics of these timings in our final paper.

Corresponding edits: Page 4, ll. 160-161.

R#2.10:

Great extensions of MRIQC to collect interval ratings and randomise the presentation.

AR: We thank the reviewer for their enthusiastic support.

R#2.11:

Could the authors add anything about the training of raters, how experienced they are or how they will be recruited?

AR: We appreciate this comment, which is very much aligned with comments 1.4, 1.5, and more precisely, 1.6, from R#1. We have carefully added details regarding the rater experience, indicating how they will be recruited, trained and their experience at the onset of the collection recorded.

Corresponding edits: Page 4-5, ll.137-143.

R#2.12:

3T imaging site only included: While I accept the reasoning to avoid filed strength as a variable, it would be possible to run a separate experiment on the 1.5T data. There are many experienced radiographers who would be able to conduct a skilled manual quality assessment of 1.5T data, and MRIQC biases would be systematic. It would be interesting to discuss/consider the possibility of repeating the protocol for the 1.5T data, given that there is so much more available ($n = 395$ over two scanners). I understand this may be outside of the scope of this project, but it could perhaps be discussed in the supplementary material as a potential follow-up.

AR: We thank the reviewer for this very relevant suggestion. It would be interesting to see whether the bias induced by defacing varies with field strength. We have added it to the discussion as a potential follow-up.

Corresponding edits: Supplementary material, Page 3-4.

R#2.13:

Experiments: - Good description of the rm-ANOVA and test of assumptions.

AR: We are glad the description is clear to the reviewer.

R#2.14:

How is the rater's confidence going to be investigated or controlled for? Or state that there will be no or only exploratory analysis.

AR: As previously discussed under comment 2.5 from this reviewer and comment 1.8 from R#1, we have clarified how the confidence data will only be used in exploratory analyses and openly shared with the publication.

Corresponding edits: Page 6, the caption of Figure 2.

R#2.15:

Visualisation: I find the BA plots difficult to read. This doesn't really impact the quality of the registration, but it would improve the reporting and publication if these were easier to engage with. e.g. shading different areas of significance, or some other visualisation.

AR: We agree with the reviewer that the BA plot needs to be modified for better readability. Importantly,

the BA plot in Figure 4 has been generated as part of our pilot study where we used categorical ratings (the latter information has been added in the manuscript following the other reviewer's remark). This led to a lot of data points overlapping. Hence, we believe that the BA plot with the interval ratings, as planned in this pre-registration, will already improve its readability. We nonetheless appreciate the reviewer's suggestions and will implement them for the final publication.

Corresponding edits: Page 8, ll. 214-217; Page 8, caption of Figure 4.

R#2.16:

Analysis of MRIQC IQMs: Please state what statistical packages will be used to perform the PCA and describe all analytic choices.

AR: We thank the reviewer for spotting this omission.

Corresponding edits: Page 9, ll. 228-230.

R#2.17:

The authors state that data will be shared CC-BY. The original IXI Dataset is shared CC-BY-SA-3.0. The CC-BY-SA-4.0 license states that any adapted or remixed data must also be shared on the same (CC-BY-SA-4.0) licence. Please confirm whether this is the case for CC-BY-SA-3.0 and update your proposed data sharing license accordingly.

AR: We thank the reviewer for calling our attention to this aspect. Most of the materials we will openly share (tables of IQMs, human assessments, training materials, etc.) cannot be considered an "Adaptation" of the original work (CC-BY-SA-3.0, sec. 1.a.), and therefore the Share-Alike clause does not apply. These new materials will be licensed under the CC-BY. However, and out of an abundance of caution, we will share the derived MRIQC *individual reports* under the (compatible) CC-BY-SA-4.0 license, meeting the requirements of the IXI dataset's license. The data and software availability statement has been updated accordingly.

Corresponding edits: Page 10, ll. 259-261.

R#2.18:

I was not able to access the code ocean capsule ("You don't have permission to access this Capsule."). Please could access be provided to my email address?

AR: We proposed the Code Ocean's capsule as a means to allow anonymous peer-review of the code. Since the reviewers signed their review and are familiar with GitHub, we have invited them (@cmorgs and @cassgvp) to access our private GitHub repository. This repository will be made public with the submission of Stage 2 of this registered report.

R#2.19:

Supplementary material - Really good discussion of the implications ("Confirming our hypothesis would build a strong support to the idea of sharing QA/QC assessments..."). I would like to see part of this included in the introduction to the main RR material. More explicitly, this is an important step to mitigate some of the impact of the necessary and appropriate pseudonymisation

processing.

AR: We acknowledge the appeal of including this discussion in the main RR material. However, following the recommendation of our editor, we removed them in our re-submission. The rationale of such a change is that we can only speculate on the impact of defacing on the quality of images perceived by humans and machines. This speculative discussion will be included in the final report, should our hypotheses be confirmed. Otherwise, we will transparently report how data disputed our original expectations about the impact of defacing on QA/QC.

R#2.20:

I would be interested to see a brief discussion of https://www.nitrc.org/projects/mri_reface on recovering/replacing some of the lost quality assessment signal.

AR: We have indicated this idea as a potential future line of work. Since the "defacing" tool of choice very likely interacts with the size of the effect, we opted for *PyDeface*, which has seen wide adoption over the years. Reversing the argument, this study could be interpreted as a first attempt to set objective criteria for the benchmarking of defacing tools. To this end, we would favor tools that maximally preserve the information content of the original image (i.e., noise distribution and potential artifacts around the face), while most effectively concealing identifying information such that it cannot be recovered.

Corresponding edits: Supplementary material, page 3.

R#2.21:

Does the research question make sense in light of the theory or applications? Is it clearly defined? Where the proposal includes hypotheses, are the hypotheses capable of answering the research question?
Yes. The theory ("defacing removes information that might be relevant for humans and for automated decision agents in their assessment of image quality") is only lightly explored. I would like to see more about the explicit features of the images which are used to raters to make a quality assessment, how they are impacted by data removal, and how they relate to the MRIQC measures. A more complete exploration of this would strengthen conclusions drawn at Stage 2.

AR: We appreciate that the reviewer found the research question reasonable and the analysis proposed adequate to address the question. We will make sure to include additional (unregistered) explorations to strengthen the conclusions at Stage 2.

Corresponding edits: Supplementary material, page 3.

R#2.22:

Is the protocol sufficiently detailed to enable replication by an expert in the field, and to close off sources of undisclosed procedural or analytic flexibility?
Yes. Missing some computational details on PCA. I was not able to access the code. I hope/assume the modified MRIQC toolbox code with manual rating widget has been shared under an appropriate licence for others to replicate the procedure.

AR: We thank the reviewer for this comment and have added computational details about PCA in the manuscript. Within the new Data and software availability section, we have clearly stated that all these developments, as well as the code to replicate the analyses will be publicly available under an Apache-2.0 license, if they are not already available under that license.

Corresponding edits: Page 9, ll. 228-230; Page 6, caption of Figure 2.

R#2.23:

Is there an exact mapping between the theory, hypotheses, sampling plan (e.g. power analysis, where applicable), preregistered statistical tests, and possible interpretations given different outcomes?
The sample size is convenience driven rather than derived from a power calculation. I would like to see this discussed as a limitation and the potential to increase the sample size by working with the 1.5T data. Statistical tests are appropriate and well-described.

AR: We agree with the reviewer that our sample size is convenience-driven, but we indeed presented power calculations. In the scarcity of prior work with reported effect sizes, we carried out a sensitivity analysis to determine the minimum effect size the statistical modeling and sample size can reliably detect. Including the 1.5T data of the IXI dataset would indeed increase the sample size, but it may not necessarily increase statistical power. However, we acknowledge the relevance of this point raised by the reviewer and have carefully revised the text describing our sensitivity analyses, and compared these estimations to previously reported results and our pilot study. We have additionally added more details on how we will interpret different outcomes of the tests. In particular, when writing the final Stage 2 report, we will be careful to clearly state that our results for manual QA may not generalize to other field strengths (e.g., 1.5T) as a limitation, and leave such analyses for future work.

Corresponding edits: Page 6, ll. 182-185; Page 9, ll. 237; Page 9, ll. 240-245; Page 13-14, Table 1; Supplementary material, pp. 3-4.

R#2.24:

For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions?
Yes.

AR: We thank the reviewers for their assessment.

R#2.25:

Is the sample size sufficient to provide informative results?
See above discussion of sampling plan.

AR: Discussion of sampling plan found under comment 2.23.

R#2.26:

Where the proposal involves statistical hypothesis testing, does the sampling plan for each hypothesis propose a realistic and well justified estimate of the effect size?

See above discussion of sampling plan. There was no clear link between effect sizes observed in the pilot (Povins, 2022) and the effect sizes expected here.

AR: We thank the reviewer for pointing out this oversight. We have edited the manuscript accordingly, in light of the discussion about the sample size justification of R#2.23.

Corresponding edits: Page 6, ll. 182-185; Page 8-9, ll.240-245; Supplementary material, page 2.

R#2.27:

Have the authors avoided the common pitfall of relying on conventional null hypothesis significance testing to conclude evidence of absence from null results? Where the authors intend to interpret a negative result as evidence that an effect is absent, have authors proposed an inferential method that is capable of drawing such a conclusion, such as Bayesian hypothesis testing or frequentist equivalence testing?
Bayesian analysis is described for human rater assessments (hypothesis 1). Could Bayes or some other evidence be used to support the MRIQC analysis? This might also be important because of the [apparent] lack of hypothesised directionality in the MRIQC IQM effects, and the high number of IQMs which will be returned by MRIQC.

AR: Although there is no explicit statement by the reviewer regarding the mentioned pitfall, we interpret that the reviewer agreed we will not conclude the absence of biases should our tests be negative. Regarding the Bayesian analysis, we have reconsidered its utilization for model comparison, as introduced previously in the response to comment R#1.13.

R#2.28:

Have the authors minimised all discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features? Maintaining this clear distinction at Stage 1 can prevent exploratory analyses at Stage 2 being inadvertently presented as pre-planned.
I have suggested that all acquired data sources should be explicitly disclosed, including the artefact descriptions obtained during the manual assessment via MRIQC, and rater confidence ratings as shown in the figures. I have suggested the authors state that there would be no planned analysis of those measures, or only exploratory. I feel acknowledging these data sources in Stage 1 is important as a matter of transparency, and is a higher priority than the risk of misreporting in Stage 2. Happy to concede to the Editor on this if my view point is not consistent with PCI recomendations.

AR: We thank the reviewer for the clarity and agree on the relevance of this point. Accordingly, potential exploratory analyses have been clearly labeled as such, following through with the recommendations previously given by the reviewer (comments 2.5 and 2.14).

Corresponding edits: Page 6, caption of Figure 2.

R#2.29:

Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done?

Yes. Provins et al. 2022 is referenced as the pilot. This work is a replication of Provins et al. 2022 on a larger dataset. There is a clear description of how the test data have al;ready been interacted with.

AR: We thank the reviewer for their enthusiastic feedback.

R#2.30:

Have the authors prespecified positive controls, manipulation checks or other data quality checks? If not, have they justified why such tests are either infeasible or unnecessary? Is the design sufficiently well controlled in all other respects?
Limited discussion of how the manipulation check of defacing will be assessed ("Only if PyDeface fails resulting in the preservation of substantial facial features from the original image, will images be excluded from the analysis", line 101- 103). Could be more explicit about how/why images will be "failed".

AR: We thank the reviewer for raising this important point. Prompted by their assessment, we have reconsidered this manipulation as its implementation could induce biases hard to identify and control for. Instead, we will not exclude any images based on the "goodness" of the defacing. Should our hypothesis be correct, images partially preserving facial features should show smaller biases on average in their rating with respect to the corresponding nondefaced condition. Therefore, excluding these images could inflate the effect size we estimate.

Corresponding edits: Page 4, ll. 111-115.

R#2.31:

When proposing positive controls or other data quality checks that rely on inferential testing, have the authors included a statistical sampling plan that is sufficient in terms of statistical power or evidential strength?
Unsure.

AR: In order to strengthen our sensitivity analyses for establishing minimal effect sizes we can detect, we have thoroughly revised the corresponding descriptions. Additionally, we have included corrections regarding the type of effect size estimations (Cohen's d, Cohen's f, and η^2). We hope these actions are found sufficiently compelling by the reviewer.

Corresponding edits: Page 6, ll.180-181; Page 7, caption of Figure 3; Page 9, ll. 238-240; Page 9, ll. 245-248; Page 12, caption of Figure 6; Page 13-14, Table 1; Supplementary material, page 2.

R#2.32:

Does the proposed research fall within established ethical norms for its field? Regardless of whether the study has received ethical approval, have the authors adequately considered any ethical risks of the research?
The authors will be using and resharing data acquired in UK NHS settings under GDPR, and that data by necessity will contain identifiable facial images at some stages of the processing. I would like to see a reference included to the authors institutional policy or governance describing how data is being handled in accordance with local data security processing guidelines, or terms

> *of use imposed by the data originator or their local restrictions. Secondary data analysis of this type does not usually require ethical approval, however it should be described in a Data Privacy [Impact] Assessment or equivalent outside of the EU. This is not yet a well understood and established norm in this field, however given that this research is specifically looking to explore the effects of privacy related processing on the data, it would be valuable to show the data are being handelled with high regard for participant privacy.*

AR: The presented study does not attempt to re-identify individuals, nor the research could be utilized to facilitate such unethical behavior. Should the IXI dataset recall the data for any of the participants following a GDPR request, we will immediately recall the MRIQC visual reports corresponding to removed participants. These two ethical aspects have clearly been stated in the manuscript.

Corresponding edits: Page 4, ll.121-125.