*Date: 02 Feb 2023*

**Review for: Defacing biases in manual and automated quality assessments of structural MRI with MRIQC**

Overview:

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.

Methodological Issues:

* 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.

For the manual quality ratings

* 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)

* 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?
* 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.
* What is the experience of the raters? If they are not known yet, will there be any requirements for the raters previous experience?
* 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)
* 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?
* How are the biases for raters (from the Bland-Altmann plots) to be used – combined in some way or just presented for each rater?

For the automated IQM analysis

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

Other

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

Minor Issues:

* Please specify what type of automatic analysis methods/steps Sitter et al found failed with defaced images (P1, L33).
* Qualify what the “very limited reliability” of automated methods is being referred to P2, L50
* 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.
* 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
* 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
* Institute of Psychiatry is now the Institute of Psychiatry, Psychology & Neuroscience
* 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
* P4 L112 – SNR at 3T is theoretically double that of 1.5T, so you could simply state that rather
* P4 L127 – effortless might be better word than “costless”
* 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

Thank you for the opportunity to review this report and all the best for the planned work.

Catherine Morgan

University of Auckland