**Extended Protocol**

**Uncovering and mitigating bias in large, automated MRI analyses of brain development**

Safia Elyounssi1,2\*, Keiko Kunitoki1,2\*, Jacqueline A. Clauss1,2, Eline Laurent1,2, Kristina A. Kane1,2, Dylan E. Hughes1,2,3, Casey E. Hopkinson1,2, Oren Bazer1,2, Rachel Freed Sussman1,2, Alysa E. Doyle1,4, Hang Lee5, Brenden Tervo-Clemmens1, Hamdi Eryilmaz1,2, Randy L. Gollub1,2, Deanna M. Barch6, Theodore D. Satterthwaite7,8,9, Kevin F. Dowling1,10, Joshua L. Roffman1,2

\*Equal authorship

1Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School

2Martinos Center for Biomedical Imaging, Massachusetts General Hospital

3Departments of Psychiatry & Biobehavioral Sciences, University of California, Los Angeles

4Center for Genomic Medicine, Massachusetts General Hospital

5Biostatistics Center, Massachusetts General Hospital and Harvard Medical School

6Department of Psychological and Brain Sciences, Washington University in St. Louis

7Department of Psychiatry, University of Pennsylvania Perelman School of Medicine

8Penn Lifespan and Neuroimaging Center, University of Pennsylvania Perelman School of Medicine

9Penn-CHOP Lifespan Brain Institute

10Department of Psychiatry, University of Pittsburgh School of Medicine

Version date: October, 2024

Corresponding author:

Joshua L. Roffman MD, MMSc

Massachusetts General Hospital

149 13th St, Room 2616

Charlestown, MA 02129

617-724-1920

[jroffman@partners.org](mailto:jroffman@partners.org)

Copyright © 2024, the authors. This document may be freely shared for academic research purposes without obtaining permission by the authors. Please cite the accompanying paper in research publications that utilize this method:

[Insert reference here]

**I. Annotated manual quality control ratings guide**

This manual quality control (MQC) ratings guide was developed based on (1) previously published methods that describe specific artifacts that are frequently encountered in structural MRI scans1-3; see also FreeSurfer wiki documentation at <https://surfer.nmr.mgh.harvard.edu/fswiki>; and (2) our lab’s extensive prior use of manual edits4, which enabled us to further classify scan quality based on the estimated time that would be needed to conduct these manual edits.

Overall assessment of each T1 volume proceeds as follows:

1. Visual inspection of full volume in all 3 planes to identify cysts, areas of (a) signal dropout, (b) large cysts (>1 cm3), or (c) large defects (ghosting, rings) that would result in systematic measurement error for cortical morphometry. These are coded separately (dropout, cyst, or “4”/unusable rating, respectively).
2. Slice-by-slice assessment in at least two planes (e.g., coronal and horizontal) for smaller-scale problems that would require manual edits. These include errant inclusion of meninges or skull tissue, intensity normalization errors, gray-white matter parcellation errors, or other abrupt changes in cortical volume that are not accounted for by slicing through gyral edges (e.g., thickness reduction that appears only in one slice would not be flagged).
3. Rating scans as follows: “1” for scans with no or only a small number of problems that would require manual edits, and that would require approximately 30 minutes for a trained technician to complete these edits; “2” for scans with a moderate number of problems that would require manual edits, and that would require approximately 1 to 2 hours for a trained technician to complete these edits; “3” for scans with a large number of problems that would require manual edits, that would require at least several hours for a trained technician to complete these edits; “4” for scans with a very large number of problems that would be impossible or impractical to edit.

The following section provides annotated examples of scans that were rated as “1”, “2”, “3”, and “4”.

**“1” scan**

A brain scan with yellow arrows

Description automatically generated

A coronal T1 slice from a volume that received a rating of “1” is shown above. This rating reflects the image’s high quality due to the absence of motion artifact, and its distinctly showing the beginning and end of white matter and pial borders. Overall, there is clear and consistent distance between the white matter and pial borders, as expected; they do not converge. The top right and two bottom arrows indicate small segmentation errors, where the pial (red) border extends into the overlying meninges. Given the high quality of the image, removing the meninges is a straightforward process as described in the FreeSurfer tutorial (<https://surfer.nmr.mgh.harvard.edu/fswiki>). Note that the arrow pointing to the white matter (yellow) border shows a small pocket extending into the gray matter. To determine whether this is an error, examining slices on either side of the current one will be helpful – as what we are seeing here may simply reflect early detection of white matter (continuous with adjacent slices). For all scans rated “1” this rate of errors is consistent throughout all slices, with some showing little to no errors.

**“2” scan**

A picture containing indoor, birthday, decorated

Description automatically generated

A coronal T1 slice from a volume that received a rating of “2” is shown above. This rating reflects the image’s moderate quality due to the presence of some degree of motion artifact (discernible rings in cortex). While white matter and pial borders should not converge, there are a few instances where they do, as indicated by the bottom right arrow pointing to the white matter (yellow) boundary that ends adjacent to the pial (red) boundary. Additionally, arrows pointing to the pial (red) border indicate a greater extension into the meninges, occurring in significantly more areas than in a scan rated “1.” Despite the motion artifact, the boundary between the pial border and meninges remains clear, facilitating edits to remove the meninges. There are also significantly more instances where the ends of the white matter border create irregular shapes without capturing any white matter. Viewing adjacent slices typically confirms that these are indeed errors. For scans rated “2”, this rate of errors is consistent throughout all slices, though some may show fewer errors.

**“3” scan**

**Graffiti on a wall

Description automatically generated**

A coronal T1 slice from a volume that received a rating of “3” is shown above. This rating typically reflects the image’s substantial artifact resulting from motion, preprocessing, or both. In this case, compared to the images rated “1” and “2”, there are significantly more errors in overestimation of white matter and pial borders, as well as instances of convergence between the two borders. For scans rated “3”, this rate of errors is evident in most slices, with significant errors present throughout.

**“4” scan**

**A picture containing invertebrate, mollusk, close, blurry

Description automatically generated**

A coronal T1 slice from a volume that received a rating of “4” is shown above. This rating reflects severe motion artifact, resulting in missing segments of the volume and blurred white matter and pial boundaries. For scans rated “4”, these characteristics are typical and cannot be corrected with manual edits.

**II. Suggested protocol for balancing type I and type II error using manual quality control ratings**

This method relies on the principle that effect size (standardized beta) and standard error of “gold standard” group (e.g., MQC=1 for manual QC) should be the most accurate – regardless of statistical significance. If, after adding in lower-quality scans, new regions emerge as statistically significant ***and*** the effect size for the included lower quality scans falls within the SE for gold-standard scans, we advise treating these regions as true positives. Such regions would have been missed in the gold-standard subsample to due insufficient power (i.e., type II error). However, if the effect size when using the included lower-quality scans *exceeds* that for the gold-standard scans (i.e., falls out of the gold-standard SE range), these should be regarded as false positives (i.e., type I error.) We suggest including surface hole number along with other technical (e.g., estimated intracranial volume), biological (e.g., sex, age, pubertal stage), and design-related (e.g., site, scanner, subject and/or family ID) covariates in analyses as appropriate.

See below for a flow chart demonstrating how to implement this method.



Alternatively, investigators may decide to remove scans with lower quality ratings (e.g., “4”, “3”) from the analysis entirely. For ABCD baseline data, “3”- and “4”-rated scans comprised only 13.2% and 2.1% of scans, respectively. Arguably the small gains in statistical power by including these scans are offset by the risk of introducing type I error (false positives) or type II error (blurring out true signal from higher quality scans).

**III. Suggested protocol for balancing type I and type II error using surface hole numbers**

This method also relies on the principle that effect size (standardized beta) and standard error of “gold standard” group should be the most accurate – regardless of statistical significance. However, if manual quality control data are not available, then the “gold standard” can be defined by an automated measure, surface hole number (SHN). Based on our analysis in the accompanying paper, a cutoff value of 29 SHN is the optimal proxy for the upper limit of MQC=“1” scans. Thus, scans with SHN≤29 are considered the “gold standard” for automated analyses (SHN Tier A), and the standard error around the standardized beta for analyses that include only SHN Tier A scans should be used as the reference range. If, after adding in lower-quality scans (SHN Tier B, C, D; see cutoffs in flow chart below), new regions emerge as statistically significant ***and*** the effect size for the included lower quality scans falls within the SE for Tier A scans, we advise treating these regions as true positives. Such regions would have been missed in the gold-standard subsample to due insufficient power (i.e., type II error). However, if the effect size when using the included lower-quality scans *exceeds* that for the gold-standard scans (i.e., falls out of the gold-standard SE range), these should be regarded as false positives (i.e., type I error.) We again suggest including surface hole number along with other technical (e.g., estimated intracranial volume), biological (e.g., sex, age, pubertal stage), and design-related (e.g., site, scanner, subject and/or family ID) covariates in analyses as appropriate.

See below for a flow chart demonstrating how to implement this method.



Alternatively, investigators may decide to omit lower-quality tiers (e.g., SHN Tiers 4, 3) for reasons described in **II** above.

**References**

1. Monereo-Sanchez, J. *et al.* Quality control strategies for brain MRI segmentation and parcellation: Practical approaches and recommendations - insights from the Maastricht study. *Neuroimage* **237**, 118174 (2021).

2. Beelen, C., Phan, T. V., Wouters, J., Ghesquiere, P. & Vandermosten, M. Investigating the Added Value of FreeSurfer's Manual Editing Procedure for the Study of the Reading Network in a Pediatric Population. *Front Hum Neurosci* **14**, 143 (2020).

3. McCarthy, C. S. *et al.* A comparison of FreeSurfer-generated data with and without manual intervention. *Front Neurosci* **9**, 379 (2015).

4. Eryilmaz, H. *et al.* Association of Prenatal Exposure to Population-Wide Folic Acid Fortification With Altered Cerebral Cortex Maturation in Youths. *JAMA Psychiatry* **75**, 918-928 (2018).