

Demonstrating quality control (QC) procedures in fMRI

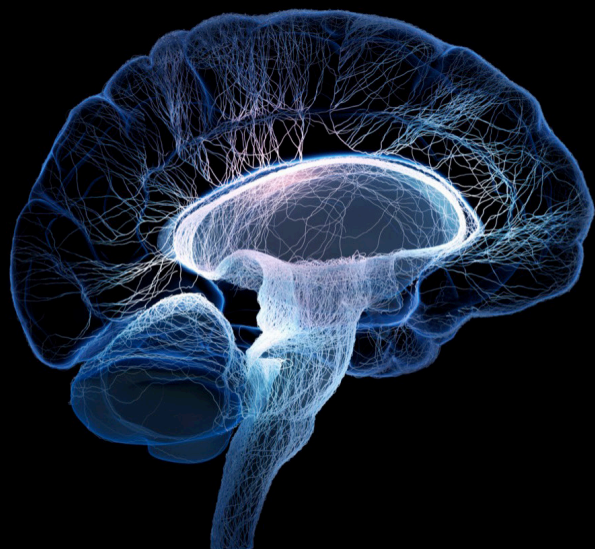
Edited by

Paul A. Taylor, Jo Etzel, Daniel R. Glen and Richard Craig Reynolds

Published in

Frontiers in Neuroscience

Frontiers in Neuroimaging



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-2704-7
DOI 10.3389/978-2-8325-2704-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Demonstrating quality control (QC) procedures in fMRI

Topic editors

Paul A. Taylor — National Institute of Mental Health (NIH), United States

Jo Etzel — Washington University in St. Louis, United States

Daniel R. Glen — National Institute of Mental Health (NIH), United States

Richard Craig Reynolds — Clinical Center (NIH), United States

Citation

Taylor, P. A., Etzel, J., Glen, D. R., Reynolds, R. C., eds. (2023). *Demonstrating quality control (QC) procedures in fMRI*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-2704-7

Table of contents

04	Editorial: Demonstrating quality control (QC) procedures in fMRI Paul A. Taylor, Daniel R. Glen, Richard C. Reynolds, Arshitha Basavaraj, Dustin Moraczewski and Joset A. Etzel
15	A functional MRI pre-processing and quality control protocol based on statistical parametric mapping (SPM) and MATLAB Xin Di and Bharat B. Biswal
27	Quality control in functional MRI studies with MRIQC and fMRIPrep Céline Provins, Eilidh MacNicol, Saren H. Seeley, Patric Hagmann and Oscar Esteban
41	Quality control practices in FMRI analysis: Philosophy, methods and examples using AFNI Richard C. Reynolds, Paul A. Taylor and Daniel R. Glen
77	Inter-rater reliability of functional MRI data quality control assessments: A standardised protocol and practical guide using pyfMRIqc Brendan Williams, Nicholas Hedger, Carolyn B. McNabb, Gabriella M. K. Rossetti and Anastasia Christakou
88	Efficient evaluation of the Open QC task fMRI dataset Joset A. Etzel
96	Demonstrating quality control procedures for fMRI in DPABI Bin Lu and Chao-Gan Yan
106	Quality control procedures and metrics for resting-state functional MRI Rasmus M. Birn
123	Functional connectivity MRI quality control procedures in CONN Francesca Morfini, Susan Whitfield-Gabrieli and Alfonso Nieto-Castañón
145	The art and science of using quality control to understand and improve fMRI data Joshua B. Teves, Javier Gonzalez-Castillo, Micah Holness, Megan Spurney, Peter A. Bandettini and Daniel A. Handwerker
161	Quality control in resting-state fMRI: the benefits of visual inspection Rebecca J. Lepping, Hung-Wen Yeh, Brent C. McPherson, Morgan G. Brucks, Mohammad Sabati, Rainer T. Karcher, William M. Brooks, Joshua D. Habiger, Vlad B. Papa and Laura E. Martin



OPEN ACCESS

EDITED AND REVIEWED BY

Matthew Brett,
University of Birmingham, United Kingdom

*CORRESPONDENCE

Paul A. Taylor
✉ neon.taylor@gmail.com;
✉ paul.taylor@nih.gov

RECEIVED 14 April 2023

ACCEPTED 12 May 2023

PUBLISHED 31 May 2023

CITATION

Taylor PA, Glen DR, Reynolds RC, Basavaraj A, Moraczewski D and Etzel JA (2023) Editorial: Demonstrating quality control (QC) procedures in fMRI. *Front. Neurosci.* 17:1205928. doi: 10.3389/fnins.2023.1205928

COPYRIGHT

© 2023 Taylor, Glen, Reynolds, Basavaraj, Moraczewski and Etzel. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Demonstrating quality control (QC) procedures in fMRI

Paul A. Taylor^{1*}, Daniel R. Glen¹, Richard C. Reynolds¹,
Arshitha Basavaraj², Dustin Moraczewski² and Joset A. Etzel³

¹Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States, ²Data Science and Sharing Team, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States, ³Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, United States

KEYWORDS

fMRI, quality control, resting state, task-based, visualization

Editorial on the Research Topic

Demonstrating quality control (QC) procedures in fMRI

Introduction

This Research Topic, “*Demonstrating quality control (QC) procedures in fMRI*”, focused on promoting quality control descriptions and discussions within the FMRI community. We invited anyone in the field to participate and perform their QC protocol of choice on sets of task-based and resting state FMRI data, describing their steps and criteria in detail. Ten teams participated, utilizing processing and QC methods that are available from a wide variety of software packages. The resulting set of articles represents a didactic resource for the field moving forward, as a reference for teaching and describing QC procedures.

The examined data collection came from real, unaltered, and publicly available datasets from widely used distributions. Even if a repository is curated, one would likely still expect to see some QC issues arise—that is one of the fundamental reasons this Research Topic was organized, and the aim of this project is certainly not to derogate the collections themselves but simply to use “real world” datasets for demonstrating detailed QC. The assortment was selected explicitly to include a full gamut of “good” to “poor” quality datasets. In the end, among the QC issues found and reported by the Project contributors were: extreme subject motion, severe ghosting, upside-down EPIs, incomplete FOV coverage, low TSNR, severe EPI distortion and dropout, left-right flipping of datasets, mismatched subjects, systematic spatio-temporal EPI artifacts, incorrect slice-timing, task-correlated motion, invalid task performance and anomalous correlation patterns. These are all issues that can affect study results, and this highlights how *anyone* undertaking an FMRI project should include careful QC assessments as part of their workflow.

Here we first describe how the focal data collection was assembled. We then give an overview of the software utilized, and highlight commonalities across the contributions of the participating teams, as well as differences and unique aspects of each. Finally, we present recommendations based on the accumulated Project contributions for the neuroimaging community around QC considerations, which apply when using either public data or acquiring one's own.

1 <https://www.frontiersin.org/research-topics/33922/demonstrating-quality-control-qc-procedures-in-fmri>

Methods

Project instructions for participants

We briefly describe the Project instructions for participants (see <https://osf.io/qaesm/> for more details). Participants were asked to perform their preferred QC steps on provided task-based and/or resting state fMRI data collections, and to describe their evaluation criteria in detail, including representative examples. Researchers could choose any desired processing steps for a whole brain study, with the final EPI data aligned to a specified MNI template (see below). “Whole brain” included subcortical structures but excluded the cerebellum, as many datasets do not fully cover the latter.

The participants could perform any QC steps they would normally use for such an analysis, using any software, visualization or processing. Each analyzed subject’s dataset would be placed into one of the following categories: “include” (passing all QC criteria; high confidence to use in a study); “exclude” (fails one or more QC criteria; high confidence to remove); and “uncertain” (questionable for whether to include).

For the Project write-ups, the participants were asked to explicitly list all their evaluative criteria, and to denote quantitative and qualitative ones. Additionally, authors should:

Describe each item listed in the QC criteria table(s) in sufficient detail for others to apply the same criteria. The criteria may also be structured as a protocol. Write the descriptions in a didactic manner, as if explaining each item to a new research assistant. Please detail quantities used.

Finally, each Project should contain a presentation of a variety of interesting and representative QC examples across each of the categories.

Dataset selection

To facilitate the QC discussions, we created a single, common collection from public repositories for participating researchers to analyze. Here we list the source datasets, as well as the approach for selecting them.

We chose to start with example investigations of commonly used data, namely human acquisitions at 3T with a single echo, which have long formed the bulk of fMRI studies. For the Project’s initial distribution of data, the acquisition site and original subject IDs were anonymized, to reduce possible evaluation biases. Since fMRI analysis is often performed on groups of subjects, and some QC factors might be considered “relative” to the group, subject ID numbering was used to identify sets of subjects from a particular site. Separate sites were labeled with group numbers, and subject IDs were simply remapped with the first digit reflecting group membership: Group 0 = sub-001, sub-002, ...; Group 1 = sub-101, sub-102, ...; etc. (see the table in the [Supplementary material](#) for the full mapping). No properties of the datasets (data values, header information, etc.) were altered in this process. The datasets are publicly available from the “fMRI Open QC Project” webpage²

(Taylor et al., 2023), which also contains further details of the Project description.

For the resting state collections, we browsed available data repositories that had open use agreements, including ABIDE-1 and ABIDE-2 (Di Martino et al., 2014), AOMIC (Snoek et al., 2021), Functional Connectome Project (FCP; Biswal et al., 2010), MPI-LEMON (Babayan et al., 2019), SALD (Wei et al., 2018), and SLIM (Liu et al., 2017), as well as a large number of OpenNeuro (Markiewicz et al., 2021) collections. In total, over 230 separate resting state data collections were initially examined for this project.

The first selection stage was to find collections with the following properties:

- Having >12 subjects, each of whom has at least one EPI and one T1w volume in the same session directory.
- EPI: TR > 1.5 s, all voxel edges < 4.1 mm, number of volumes > 100, non-zero srow values in the NIFTI header.³
- T1w: all voxel edges < 2.1 mm, non-zero srow values in the NIFTI header.

This reduced the number of collections to 56.

Then, quick processing and brief visual investigation were performed. Data collections with systemic issues, such as overly tight FOV (cutting off the cerebellar cortex), very poor EPI tissue contrast, obvious ghosting in the EPIs, and odd coordinate systems (e.g., not approximately centered around the coordinate origin, suggesting possible DICOM conversion and header issues) were removed from further consideration. From the remaining sets, we selected collections with a variety of voxel sizes, run lengths and numbers of runs, and particularly those that appeared to contain both reasonable data and a variety of occasional (but not systemic) QC considerations. To finalize the Project data collection size, we aimed to balance the breadth of data properties to explore with the number of researchers likely to participate: having more sites/subjects would likely increase the former but decrease the latter.

Therefore, we settled on having seven resting state fMRI sites from various data repositories and formed “groups” of ~20 subjects each. Most of the Project groups were subsets of their original repository collections; the subsets generally had a range of subject motion and other underlying considerations. Some repositories originally contained explicit categorization of subjects as “control” and non-control, such as having TBI (traumatic brain injury) or psychiatric diagnosis; those designations did not influence data selection, and subjects were typically drawn from multiple categorizations, as most MRI studies contain such combinations. The final list of included resting state datasets (Groups 1–7) is provided in [Table 1](#), with a brief description of properties by site/group.

Similar considerations to the above were used for selecting task-based fMRI data. As an additional factor, there are a wide variety of possible task designs, with differing degrees of complexity for modeling and analysis. Quality control considerations of the paradigm timing, both in terms of setup and subject response, are

² doi: 10.17605/OSF.IO/QAESM: <https://osf.io/qaesm/>.

³ That is, have a defined voxel grid, where the affine sform matrix in the NIFTI header is nonzero.

TABLE 1 List of the sites from which project datasets were selected, along with brief descriptions of EPI properties.

Brief descriptions of the resting state datasets used in the project
Group 1: ABIDE-1, KKI (Barber et al., 2012; Nebel et al., 2014), $N = 20$ subjects used (of 55 total). <i>FMRI acquisition details:</i> Philips Achieva 3T scanner, EPI axial slice acquisition with fat saturation and SENSE (factor=3), flip angle = 75° , TE = 30 ms, TR = 2.5 s, voxel size = $2.67 \times 2.67 \times 3.0$ mm, slice timing provided in JSON sidecar, PE direction = j-; subjects instructed to focus on a crosshair on black computer screen.
Group 2: ABIDE-1, Trinity (Delmonte et al., 2012), $N = 20$ subjects used (of 49 total). <i>FMRI acquisition details:</i> Philips Achieva 3T scanner, EPI axial slice acquisition with fat saturation and SENSE (factor=2), flip angle = 90° , TE = 28 ms, TR = 2.0 s, voxel size = $3.0 \times 3.0 \times 3.841$ mm, slice timing provided in JSON sidecar, PE direction = j-, subjects instructed to close eyes during scan.
Group 3: ABIDE-2, KUL-3 (Bernaerts et al., 2016), $N = 16$ subjects used (of 28 total). <i>FMRI acquisition details:</i> Philips Achieva Ds 3T scanner, EPI axial slice acquisition with fat saturation and with SENSE (factor=2), flip angle = 90° , TE = 30 ms, TR = 2.5 s, voxel size = $1.562 \times 1.562 \times 3.1$ mm, slice timing provided in JSON sidecar, PE direction = j-, subjects instructed to focus on a white fixation cross on black background.
Group 4: FCP, Baltimore (Pekar and Mostofsky, 2010), $N = 23$ subjects used (of 23 total). <i>FMRI acquisition details:</i> 3T scanner (unspecified type), TR = 2.5 s, voxel size = $2.667 \times 2.667 \times 3.0$ mm, subjects instructed to keep eyes open and fixate (target unspecified) during scan.
Group 5: OpenNeuro, ds000220 (Roy et al., 2017), $N = 20$ subjects used (of 26 total). <i>FMRI acquisition details:</i> Philips Achieva and Siemens Trio 3T scanners, EPI axial slice acquisition with segmented k-space (no SENSE), flip angle = 90° , TE = 34 ms, TR = 2 s, voxel size = $1.85 \times 1.85 \times 4.0$ mm, instructions to subjects undescribed.
Group 6: OpenNeuro, ds000243 (Petersen et al., 2018), $N = 20$ subjects used (of 120 total). <i>FMRI acquisition details:</i> Siemens Magnetom Trio 3T scanner, 12 channel head coil, flip angle = 90° , TE = 34 ms, TR = 2.5 s, voxel size = $4.0 \times 4.0 \times 4.0$ mm, instructions to subjects undescribed.
Group 7: OpenNeuro, ds000245 (Yoneyama et al., 2018), $N = 20$ subjects used (of 45 total). <i>FMRI acquisition details:</i> Siemens Verio 3T scanner, 12 channel head coil, flip angle = 80° , TE = 30 ms, TR = 2.5 s, voxel size = $3.0 \times 3.0 \times 3.51$ mm, slice timing provided in JSON sidecar, subjects instructed to close eyes during scan.
Brief description of the task-based state datasets used in the Project
Group 0: OpenNeuro, ds000030, “task-pamenc” (Poldrack et al., 2016; Bilder et al., 2018), $N = 30$ subjects used (of 272 total). <i>FMRI acquisition details:</i> Siemens TrioTim 3T scanner, EPI acquisition with segmented k-space and fat saturation (acceleration factor PE = 2), flip angle = 90° , TE = 30 ms, TR = 2 s, slice timing provided in JSON sidecar, PE direction = j-.

See the [Supplementary material](#) for a detailed subject list from each site. In some cases, properties varied across the site, which was noted within some QC evaluations, and properties shown here are those for the first subject in each group. Group 4/s details were not provided in original project downloads, because the dataset JSON sidecar files did not contain acquisition information. This description comes from the “Release Table (April 6, 2012)” spreadsheet from the FCP download website: https://www.nitrc.org/docman/?group_id=296. For Groups 5–7, voxel size was included only in the NIFTI dataset, not included in the JSON sidecar.

important in much of FMRI research. For this Project we decided to use task FMRI data from a single site and paradigm, and we wanted to select a relatively straightforward design with a small number of stimulus classes, to simplify explication, processing and modeling considerations.

Thirty subjects from the following task-based dataset were selected. Table 1 provides a brief description of this “Group 0,” including FMRI acquisition properties contained within the JSON sidecar files in the Project download. The specific task was a paired memory encoding task (“pamenc”) with button-pushing responses (see Poldrack et al., 2016, for details). In addition to the originally distributed events TSV file, we also provided a simplified task file with only three columns: stimulus onset time, duration and a trial type label (“TASK,” “CONTROL”). Teams were free to use either set of timing information—or even to not use any—as part of their QC. Onset timing was essentially identical for all but two subjects (whose onsets were uniformly 2 s later), separated by 2.5–18.5 s (mean = 7.5 s). Response times, which could represent event duration, had per-subject means of 0.51–1.57 s (range = 0.0–2.43 s) for CONTROL events and 0.45–2.65 s (range = 0.0–4.0 s) for TASK events. Inter-stimulus interval times ranged from 1.3 to 17.3 s (mean = 6.4 s).

We note that de-identifying the task data to fully blind teams from the source dataset was challenging, because BIDS (Brain Imaging Data Structure specification) encodes the task label explicitly in the dataset filenames. For example, an EPI dataset is called sub-001_task-pamenc_bold.nii.gz, where “pamenc” is the label for the specific task; searching online for “fmri pamenc” leads

to the original repository. Because we did not want to change any dataset properties besides the subject IDs (to avoid introducing any errors by mistake), we neither relabeled columns within the subject timing files nor changed the task label in the filenames. Therefore, in theory, participating teams could have investigated more background details about the task data; we are unaware if any did, but, in practice, essentially the same QC considerations would still apply.

In the end, the available Project data collection was comprised of seven groups of 139 total resting state FMRI subjects and 1 group of 30 task-based FMRI subjects. Each subject had one T1w anatomical reference and 1 or 2 EPI functional runs from a single session directory. These collections were intended to provide a basis for QC examples, with a full spectrum of data quality within each group and a diverse assortment of items to discuss across the subjects: having a mix of both reasonable and poor quality data would facilitate clearer depictions of QC procedures and contrasts. The collections were initially investigated for this purpose, using a quick inspection. However, during the course of the analysis for the Project itself, it became apparent with a more complete QC procedure that the EPI datasets for two groups actually did contain systemic artifacts (see Reynolds et al.). While this is certainly worth examining and understanding from a QC point of view, it had not been the intention to include such datasets within this project. This occurrence does primarily highlight two important points: (1) an in-depth quality control investigation is necessary on at least some subset of a data collection to truly understand its contents, whether using shared or acquired data; and (2) QC must be performed from

the start of data acquisition (also using an in-depth examination), to avoid the propagation of systemic issues.

The repositories from which subjects were drawn contained a wide range of age spans: from 8–13 to 56–78 years. Neither age nor sex nor any other subject-specific information was included in the accompanying participants.tsv file, as part of anonymization. In “real” fMRI studies that use a standard template to define a common final space, it is generally considered preferable to match that template to the age of the subjects, such as the Haskins pediatric template (Molfese et al., 2021) for studies of children; and, increasingly, templates and atlases exist for a wider variety of geographical locales, such as Korean (Lee et al., 2005), Chinese (Tang et al., 2010), and Indian (Holla et al., 2020) populations, which may also provide a better reference. However, since the present project was focused on subject-level QC considerations and not on a group-level report, researchers were asked to use just a single reference template for simplicity and uniformity: the widely used MNI-2009c ICBM152 T1w, non-linear asymmetric volume (Fonov et al., 2011). Any particularly notable mismatches to the template dataset would be deserving items for QC commentary by the participating teams.

BIDS packaging

The selected datasets were then merged into BIDS-valid resting state and task-based collections. We used multiple versions of the BIDS validator (1.2.5 and 1.9.9) to ensure BIDS compliance. As noted above, we did not alter the data or metadata supplied from the source dataset. Since each of the datasets was already available publicly in a BIDS structure, we only needed to rename the directories and files according to our site-based enumeration (see [Supplementary material](#)).

We first merged the seven resting state groups into a single data collection, and then deposited the appropriate top-level text files (dataset_description.json, participants.*, etc.) into each of the resting state and task-based collections. For resting state Group 4, we noticed that the JSON sidecar for the functional image in the source dataset was provided at the dataset level instead of at the participant level. To maintain consistency with the other groups, we copied this sidecar to the latter and renamed the file accordingly. We also note that for resting state Groups 4 and 5, JSON sidecars for the T1w images had not been supplied in the source dataset. Since metadata fields contained in these sidecars are often contingent on conversion software version, we opted to preserve the absence of this metadata.

We found no validation errors in the resting state collection and noted five warnings: (1) some images were not supplied with slice timing info; (2) not all subjects contained the same number of EPI files (e.g., some subjects in Group 6 had two functional runs, while the rest of that group and all other groups only contained one per subject); (3) not all subjects/sessions/runs had the same scanning parameters, sometimes even within a single group/site; (4) NIFTI header fields for unit dimensions were missing in the anatomical volumes for some subjects (xyzt_units was 0 for most of Group 1 and all of Group 2); and (5) two subjects (sub-506 and sub-507) had a mismatch between the number of items in the SliceTiming

array and the k dimension of the corresponding NIFTI volume. For the task-based collection we found no validation errors and one warning: the tabular file contained custom columns not described in the data dictionary for the timing files. We avoided altering any of these warnings, as they existed in the original data, and left these as possible QC items for teams to discuss.

Participating teams and software utilized

One goal of this Research Topic was to have as wide a representation of software tools and research labs as possible, in order to have a maximal breadth of QC descriptions. The Research Topic was advertised widely on general MRI analysis message boards, such as the INCF's Neurostars, and on email lists, such as the open “niQC” email group, which was created to foster discussions on neuroimaging quality control. It was advertised at major neuroimaging conferences and workshops, such as ISMRM and OHBM. Email notices were also sent to members of software development groups, to project consortia (e.g., ENIGMA) and to many fMRI labs across the field. In the end, there were 10 participating teams, from labs across three continents.

Across the contributions, there was a wide array of software used for each of the processing and QC phases. We list the processing and QC software packages used by each team in [Table 2](#). We note that virtually all of the tools and implemented procedures exist in freely available (and mostly open source) software. As a result, this means that this set of Topic contributions assembles detailed QC descriptions across many widely used software packages that can immediately be used across the field for training, processing and research applications.

Results

Common themes across teams

There were several common themes running across the participating teams' analyses.

1) Each team found subjects to exclude based on one or more aspects of data quality. As noted above, these collections all come from standard public data repositories. These repositories are great resources for the field for open data sharing, increasing multisite studies and having validation datasets, but there should generally *not* be the expectation that they are fully curated for data quality (and as noted in below in Theme 7, it may be impracticable to do so in a general way). Exclude-or-uncertain rating fractions varied across the teams, but many excluded 25% or more ([Figure 1](#)). In some cases, subtle but systematic artifacts were even found that led to recommending the complete exclusion of Groups 2 and 4 (see [Reynolds et al.](#)). These findings stress the importance of performing QC: *researchers should always check that data contents are appropriate for their study, whether acquiring collections themselves or downloading them.*

2) Each team evaluated one or more subject's datasets as “uncertain.” This is reasonable and expected, particularly when first investigating a data collection. This categorization would almost by definition be expected to be heterogeneous

TABLE 2 Software used by each participating team for data processing and quality control.

Team	Software for processing	Software for QC
(A) Birn	AFNI, FSL, ANTs	AFNI
(B) Di and Biswal	SPM, Matlab	SPM, Matlab
(C) Etzel	fMRIPrep (with ANTs, AFNI, FreeSurfer, FSL, Nipype)	R (with knitr, RNIfti and fields), AFNI
(D) Lepping et al.	AFNI	AFNI, REDCap
(E) Lu and Yan	DPABI, DPABISurf, DPARSF, fMRIPrep, FreeSurfer, ANTs, FSL, AFNI, SPM, PALM, Matlab, DARTEL	DPABISurf, DPARSF, fMRIPrep, Matlab
(F) Morfini et al.	CONN (with ART), SPM12, Matlab	CONN, SPM12, Matlab, FSLeys
(G) Provins et al.	MRICQC (with ANTs, AFNI, FreeSurfer, FSL, Nipype, SynthStrip), fMRIPrep (with ANTs, AFNI, FreeSurfer, FSL, Nipype)	MRICQC (with ANTs, AFNI, FreeSurfer, FSL, Nipype, SynthStrip), fMRIPrep (with ANTs, AFNI, FreeSurfer, FSL, Nipype)
(H) Reynolds et al.	AFNI, FreeSurfer	AFNI
(I) Teves et al.	FreeSurfer, AFNI	AFNI
(J) Williams et al.	FSL, cinnqc (with FSL and pyfMRIqc)	pyfMRIqc

Citations for each are included here, in alphabetical order: AFNI (Cox, 1996), ANTs (Avants et al., 2012), ART (Ardekani and Bachman, 2009), cinnqc (<https://github.com/bwilliams96/cinnqc>), CONN (Whitfield-Gabrieli and Nieto-Castanon, 2012; Nieto-Castanon, 2020), DARTEL (Goto et al., 2013), DPABI (Yan et al., 2016), DPABISurf (Yan et al., 2021), DPARSF (Yan and Zang, 2010), fields (Nychka et al., 2017), fMRIPrep (Esteban et al., 2019), FreeSurfer (Fischl and Dale, 2000), FSL (Smith et al., 2004), FSLeys (McCarthy, 2022), knitr (Xie, 2014), Matlab (www.mathworks.com), MRICQC (Esteban et al., 2017), Nipype (Gorgolewski et al., 2011), PALM (Winkler et al., 2014), pyfMRIqc (Williams and Lindner, 2020), REDCap (Harris et al., 2009, 2019), RNIfti (Clayden et al., 2020), SPM (<https://www.fil.ion.ucl.ac.uk/spm/>; Ashburner, 2012), and SynthStrip (Hoopes et al., 2022).

across researchers, given their different backgrounds, experience, opinions, expectations and intended use for the data. QC considerations and criteria will adapt over time, likely reducing the number of uncertain evaluations, but it is still a useful categorization to have in a QC procedure. It is essential to identify unknown or “surprising” features of a data collection or processing procedure. In a real study, this rating would likely be a temporary evaluation that leads to checking acquisition or other aspects more in-depth, perhaps even leading to a corrective measure or change in the acquisition. A subject given this rating may eventually be evaluated as either include or exclude.

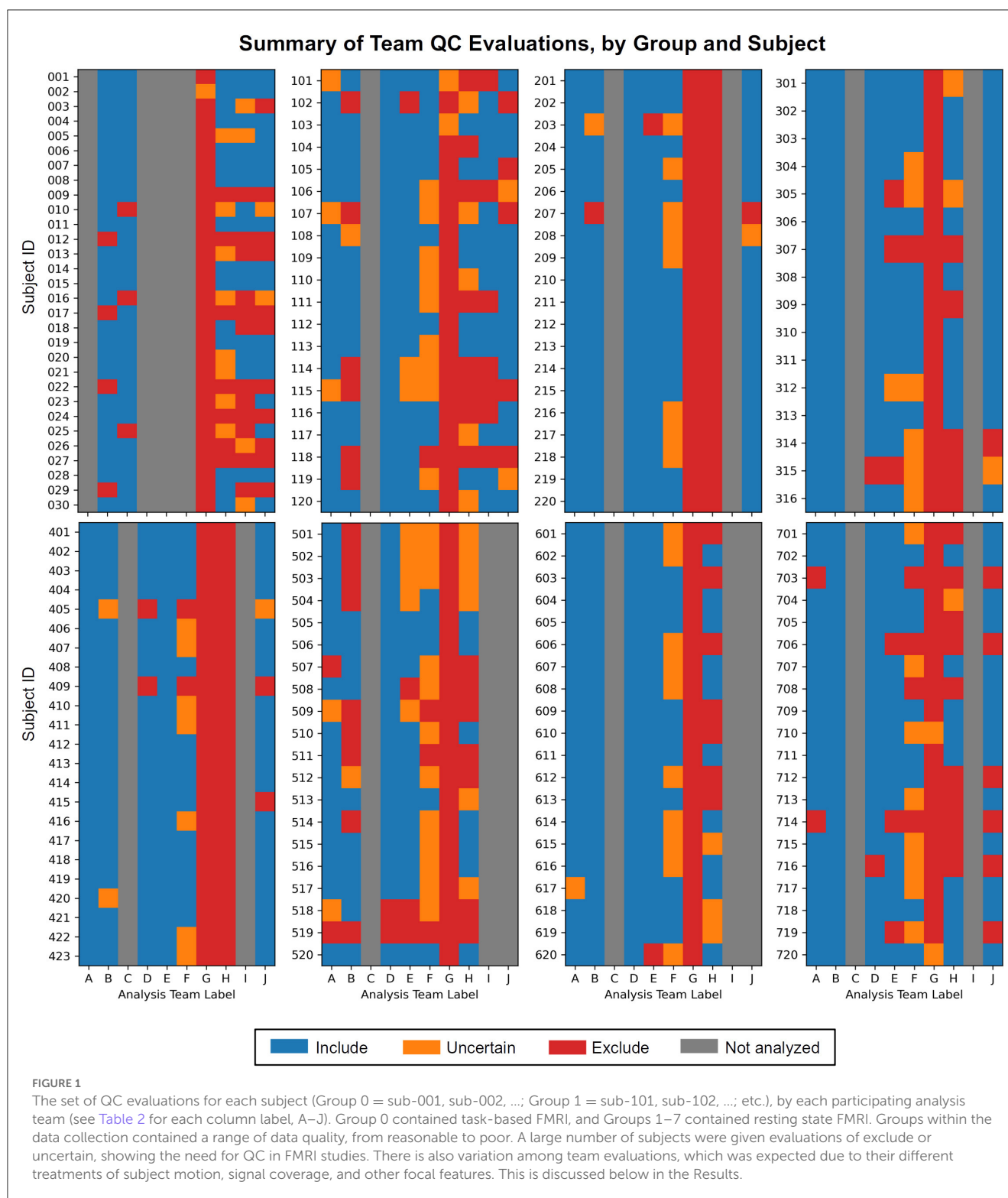
3) Nearly all QC protocols started by investigating the unprocessed data’s consistency and “metadata” properties. These included checking the number of EPI runs, voxel sizes, acquisition parameters, and other properties that are generally contained in the NIFTI headers and/or JSON sidecars; standard data collections are also likely to be accompanied by further subject descriptions (age, demographic, etc.). Even when acquiring one’s own data, it is necessary to be sure that these underlying properties are consistent and meet expectations. Alterations in scanner settings, software version, DICOM field conversion and more can easily occur, and these can detrimentally alter dataset features, affecting final results or compatibility for inclusion within a study.

4) Each team identified consistency, reliability or mismatch errors within subject datasets. For example, all teams found two datasets that had upside-down EPI datasets, and some also identified left-right flip errors between a subject’s EPI and anatomical volumes, which is a disturbingly common problem in fMRI (see Glen et al., 2020). Two teams even suspected that a subject’s EPI and anatomical volumes came from different subjects, based on sulcal and gyral pattern mismatch. These kinds of fundamental data issues are difficult, if not impossible, to reliably correct after the fact. Some

groups chose to address the EPI-anatomical consistency issues by assuming the anatomical dataset was correct, but while that may produce EPI-anatomical consistency, the presence of such header problems greatly reduces the reliability in absolute left-right identification. As was noted by multiple teams, fiducial markers are needed for clear identification (and some were identified in the visual inspections of some Project datasets here).

5) Each QC protocol used qualitative criteria and visual inspection of datasets. These included checking the raw data and inspecting derived images (e.g., TSNR or standard deviation maps) for suitability, as well as for artifacts. Visual checks were also used to evaluate the success of processing steps, such as alignment or statistical modeling. While these procedures cannot be performed automatically, they benefit greatly from systematization within a QC protocol, which software developers aim to facilitate. These qualitative checks carry the requirement for researchers to learn how to distinguish reasonable and problematic data, as well as to accurately communicate their methodology.

6) Most, but not all, protocols included quantitative/automatic checks. The teams employed a variety of quantities based on subject motion, TSNR and other measures. These tests are useful and find some of the most common kinds of expected problems. It was perhaps surprising that not every protocol included quantitative checks (while all *did* include qualitative ones, noted in Theme 5). This may reflect that visualization is still key to evaluating several data features and processing steps, and quantitative criteria typically originate as useful extensions of such understanding. It is likely that more developments for automating certain checks will be made over time, but this process also typically is rooted in visual inspection during the “training phase” of determining meaningful quantities and reasonable thresholds.



7) QC parameters were closely tied to a specific analysis and research goal. Nearly every group made the point that some subject data and data collections may be appropriate for one particular analysis but not another. As a consequence, it is likely not possible to simply adopt existing QC ratings on a given data collection from a separate study when using that data for a new project. While

prior QC evaluations may inform those of a new analysis, the burden is always on the researcher to be sure of the contents of the data for their current application. There is no “one size fits all” set of criteria, as there is no single method for designing a study (sample size, number of groups, task paradigm, etc.), acquiring data (different field strengths, echo number, etc.), performing

analysis (ROI- vs. voxel-based; surface vs. volumetric; etc.) and so on.

8) Non-EPI items can affect fMRI analysis, too. While the vast majority of QC evaluations focus on the EPI volume and its spatiotemporal properties, checks on the accompanying data can also affect the usability of the dataset as a whole. For example, some cases of notable anatomical variability were cited by most teams, such as having extremely large ventricles (and its limiting effect on the accuracy of template registration), as well as other anatomical anomalies. In other data collections one might find alterations to structure due to tumors, surgery or hemorrhages, which might necessitate removing a subject from the analysis or at least constrain the analysis options. Similarly, evaluating the stimulus timing in its own right was shown to be useful (Etzel; Reynolds et al.). For more complicated study designs, one might also analyze accompanying data such as physiological time series (such as cardiac and respiratory rate), etc. All the input data used for analysis needs to be reviewed.

9) Each team made their processing and QC pipelines publicly available. This kind of open processing (e.g., using GitHub, OSF or another accessible webpage) is becoming more common within the field, but it is important for this practice to keep growing. Given the didactic nature of this Research Topic, we hope that having these methods directly available will encourage the implementation for more detailed QC protocols and reporting.

Individual elements and focuses among teams

Each of the submissions also introduced their own unique perspective and tools for quality control. We briefly list some examples here.

Birn analyzed the seven resting state groups. This paper explored the effects of using different motion thresholds, as well as the inclusion/exclusion of low-frequency fluctuation bandpassing, during processing. In particular, it investigated some trade-offs of trying to remove artifactual features with reducing the degrees of freedom in each subject's data, using network based dissimilarity matrices of QC-FC (Ciric et al., 2017; see below) that can be used for quality control evaluation.

Di and Biswal analyzed 1 task group (using stimulus timing) and the seven resting state groups. The authors included tissue-based segmentation estimates within their visual checks of anatomical-to-template volumetric registration. Tissue-masks were also used within a set of time-series checks of subject motion-related artifacts, where principal components of white matter and cerebrospinal fluid masks were examined for similarity with motion regressors and global mean signals.

Etzel analyzed 1 task group (using stimulus timing). This work focused on the task-based fMRI data. Among other QC steps, it included checks for participant behavior and responsiveness, such as by basing some criteria on patterns of button-pushing. Being confident that subjects had followed the task assignment to a reasonable degree is indeed paramount in neuroimaging, whether

for explicit task-based paradigms or for naturalistic and resting state ones (with eye-tracking, alertness monitoring, etc.).

Lepping et al. analyzed the seven resting state groups. While all teams had an explicit list of QC criteria, this team created a REDCap checklist form to itemize and store the dataset assessments. They emphasized how this system facilitated the training and replicability aspects of the QC, which are vital aspects in any evaluation procedure. This also provided a convenient mechanism for sharing QC results.

Lu and Yan analyzed the seven resting state groups. This team included surface-based processing and criteria as part of their QC procedure, even though the analysis itself was explicitly volumetric. This allowed the evaluation to contain an interesting intersection of anatomy- and function-based assessment. They also briefly explored the differences of group-level tests with and without incorporating their excluded subjects.

Morfini et al. analyzed the seven resting state groups. Among other QC criteria, this team used multiple "QC-FC" analyses (Ciric et al., 2017) to evaluate the data at the group-level, an approach which incorporates both the quality of underlying data itself and the denoising/processing steps utilized on it. For example, one QC-FC measure involved calculating correlation matrices from 1,000 random voxels across a gray matter template in standard space.

Provins et al. analyzed one task group (not using stimulus timing) and the seven resting state groups. This work included exclusively qualitative assessments of quality, including signal leakage from eye movements, carpet plots and ICA-based components. One particular point of emphasis was on the importance of examining "background" features within the field of view (FOV), as patterns there can reveal several kinds of artifacts, such as aliasing ghosts, subject motion spikes, or scanner issues.

Reynolds et al. analyzed one task group (using stimulus timing) and the seven resting state groups. This QC procedure was organized into 4 or 5 separate stages for the resting state and task fMRI data, respectively, including GUI-based checks with InstaCorr (interactive "instant correlation;" Song et al., 2017) to follow up on observed spatio-temporal features as necessary. The authors also explicitly placed QC within the larger context of understanding the contents of the dataset and having confidence in what goes into the final analysis, rather than viewing it simply as a subject selection/rejection filter.

Teves et al. analyzed one task group (using stimulus timing) and one resting state group. This team organized their work as a QC assessment guide for both new and experienced researchers, and they emphasized the importance of interactive training and discussions with new researchers. They also used the comparison of EPI-anatomical alignment cost function values as a measure to trigger a visual check for potentially mismatched datasets.

Williams et al. analyzed one task group (not using stimulus timing) and five resting state groups. In particular, this work focused on the issue of inter-rater variability and reliability. Even within a single lab performing QC, there can be different assessments of datasets: qualitative evaluations can vary, as well as the choice of specific quantitative thresholds. This issue is also critical for describing QC procedures as accurately as possible to others when reporting results.

General differences in team perspectives

Overall, there were some general differences in teams' approaches and scopes, which influenced QC discussions and selections. These did not reflect decisions that would necessarily be described as either right or wrong, but rather different choices made by teams that would contribute to variability of dataset evaluations (see Figure 1).

Firstly, the range of QC items necessarily depended on what processing steps were implemented, and the latter choice can vary widely across the field of fMRI analysis. There is no generally defined set of processing steps to apply when performing QC of an fMRI dataset. For example, some groups included subject-level (or "first level") regression modeling within their processing, while others did not.⁴

Secondly, for the task fMRI dataset, some teams chose to ignore the timing files in their QC processing, while others included the stimulus information. Some even analyzed and interpreted the performance information in detail on its own within the list of exclusion criteria (e.g., Etzel). These again reflect different choices and degrees for understanding the presented data, and will necessarily contribute to variability in subject selections. For more complicated task designs (which certainly exist within the field), one would expect the QC approaches to have further variability, and to be closely tied to the analyses at hand, such as which stimulus contrasts are particularly central to the analysis.

Thirdly, the issue of subject motion was viewed and treated differently. Some teams used estimated motion-based parameters (e.g., Enorm or FD) for censoring (or "scrubbing") time points to remove potentially contaminated volumes, and then to include the number of censored volumes within subject exclusion criteria. Other teams adopted processing approaches to mitigate effects of subject motion in other ways (within minimal or no censoring), with the stated aim of avoiding potential biases, arbitrary thresholds and loss of subject data. These philosophical choices will result in very different criteria for QC evaluations, given that typical data collections contain a range of subject motion profiles.

Fourthly, there were also different interpretations of how much signal dropout and distortion within a volume was acceptable before excluding a subject. For example, one team excluded 166 out of 169 subjects (and listed the remaining 3 as uncertain) from the evaluations of these features (Provins et al.). In a real study, this consideration might take the form of listing brain regions of particular interest and verifying the signal quality there specifically.

Additionally, beyond the fact that researchers make their own choices when determining what data are satisfactory to include in their research, the Project guidelines omitted details such as research goals, which might imply anatomical regions of particular interest. Similarly, subject group types were omitted, which might identify subjects for whom elevated levels of typical motion, or anatomical anomalies, would be expected. Researchers also made independent choices on how to treat within-group inhomogeneity

of acquisition, such as whether subjects were required to be scanned on the exact same grids or to have the same number of EPI runs. As such, for this Project variance in QC perspectives was expected.

Discussion

The immediate goals of this project were:

1. **To promote the broader adoption of quality control practices in the fMRI field.** There are many QC tools and protocols available in publicly available software (e.g., those in AFNI, CONN, DPARSF, fMRIPrep, MRIQC, pyfMRIqc, and SPM were all used here), perhaps more than people have typically realized, and this set of Research Topic articles provides a didactic collection of them for researchers and trainees to use.
2. **To facilitate the inclusion of more details in QC protocol descriptions.** Each Project contribution contained an explicit list of QC criteria, along with demonstrations of most features. We hope these help start to systematize QC reporting within the field.
3. **To develop the view of QC as more than "just" vetting datasets, but rather as more deeply understanding the contents of the collection and analysis as a whole.** This should allow for greater confidence in final results, and hopefully improve reproducibility and reliability across the field.
4. **To share QC criteria across researchers who are performing analysis and developers creating tools, thereby improving the set of available QC tools in the process.** We would expect increasing clarity and potentially homogeneity of QC methods as a result of this work.

One longer term goal is to motivate the development of new QC methods, techniques and criteria. As noted in several Project papers, MRI acquisitions and analyses are complex and always changing, so evaluation criteria should continue to adapt. For instance, new images may summarize a feature in a clearer way, or more quantitative methods could be developed to streamline QC procedures. It is our experience as methods researchers and software developers that these kinds of advances are often rooted in visualization and understanding: *quantitative checks are essentially extensions of qualitative ones, in which understanding is rooted*. The present project collected a large number of datasets with varied properties, so that many people could view and comment on them in detail—we hope this provides a useful incubator for further QC development, which can be expanded across more data collections and researchers.

Another long term goal of this project is to facilitate the development of a common language and clear description of QC items. Several teams noted that there is not currently any general commonality in criteria or descriptions in the field, and that developing one would improve the ability to use, understand and communicate QC in work. For example, even referring to an apparently straightforward mathematical measure like TSNR (temporal signal-to-noise ratio) can lead to confusion, since there can be multiple reasonable definitions. Therefore, analysts should specify which definition they are using (as well as ensure that they are using a reasonable one), not only a numerical threshold.

⁴ In some cases, this may reflect a differentiation of "processing" vs "preprocessing," in which the former includes subject-level regression while the latter does not. However, these terms are not used consistently across the field. Within the Project description, "processing" was consistently used.

QC recommendations for researchers

The following are recommendations for implementing quality control in fMRI studies, drawing from the accumulated Project contributions, guidance and suggestions.

1) Check new acquisitions immediately—delays can lead to wasted data. Performing in-depth evaluations is crucial with the first few subjects in a protocol, to avoid systematic errors from the start. Maintaining checks remains important as scanner settings can easily be changed by accident or through an upgrade, etc., and this can flag alterations in data quality or properties. Acquiring good data is always better than trying to fix problematic data retrospectively.

2) Conduct detailed QC checks whenever using a new data collection or starting a new project. Most public repositories explicitly note that curation should not be assumed, and prior checks may have focused on different purposes, regions of interest or type of analysis. QC also integrates directly with verifying processing steps, and different analyses may have different properties and requirements.

3) Treat QC as understanding data, not just “removing bad data.” fMRI datasets are complicated, and many small details can affect downstream results. Treating QC as purely the elimination of bad data can lead to selection bias and to missing systematic issues—often checking *why* some datasets get removed provides useful insight into the entire collection. Understanding the full properties (and realistic limitations) of data will generally lead to better interpretation of it. Researchers should be confident in their data and its contents, and in-depth QC is the only way to achieve this.

4) Apply both qualitative and quantitative checks.

Visual verifications remain fundamental in data analysis, as shown by the participating teams here. These can be usefully systematized for maximal efficiency and utility, and these inform and complement automatic checks of derived quantities. This combination also typically helps with the development of new QC measures.

5) Clearly define and describe all QC steps and measures. This is necessary to maintain consistency of the QC within a lab or group setting, as evaluations of features can change over time or differ among people. All quantities should be clearly defined, since there may be multiple derivations; thresholds are not useful if their associated quantities are not clearly described. Having clear checklists facilitates implementing the QC, as well as reporting it in papers and presentations.

6) Coordinate QC evaluations with the paradigm and aims of the current study. In practice, it is difficult to make one QC evaluation apply to all possible purposes, due to the variability of study design, regions of interest, etc. Viewing previous QC evaluations might be useful, but those could be missing important characteristics for the present work or be overly harsh/lenient. Include explicit QC discussions in the planning stages of each study design.

7) Ensure (at least some) in-depth QC, even for large studies. The typical amount of time, expense and per-researcher effort of acquiring any subject is large (e.g., planning, piloting, grant writing, training, acquiring, and analyzing). As many QC steps are already integrated into analysis software, the relative effort of checking data and processing quality is actually quite small compared to

that of the other stages of acquisition and analysis—QC *should not be skipped simply because it comes near the end of processing*. Big data can still be corrupted by systematic issues in acquisition and analysis. Even when applying automatic checks across all subjects, in-depth QC (including visualization and qualitative checks) should still be applied to at least meaningful subsamples across scanners and systematically across time, to avoid wasted data and resources as well as artifactual results.

8) Share QC advice and recommendations. Stating what QC steps are most useful for identifying certain features or for validating data for certain analyses benefits everyone in the neuroimaging community. Similarly, adding new tests and features helps other researchers and software developers directly.

9) Make QC scripts public where possible. While textual descriptions of methods in papers are useful and provide explanatory context, there are many influential details for both processing and QC that exist only at the level of code. Researchers presenting findings in posters, talks and other presentations would also be encouraged to provide links to their processing and QC scripts. Having the code available provides a valuable resource for the field, and hopefully this will help promote the wider adoption of QC integration into fMRI processing.

10) Make QC evaluations public where possible. Many of the QC protocols and software tools implemented in Project contributions produced reports that can be shared and/or archived. These include PDFs, HTML pages, RedCap reports, and JSON files. These could be included in NeuroVault uploads, for instance, as well as linked to papers, presentations and data repositories. Additionally, provide QC feedback to public repository hosting sites and/or to the researchers who acquired the original data: just like software packages, data collections have version numbers because fixes and updates can be required; QC feedback can benefit the neuroimaging community.

11) Stay up to date with QC developments. QC measures and methods will change over time. New acquisition and analysis approaches will lead to new artifacts and other considerations to evaluate; new ideas and software developments provide new checks and solutions.

Conclusions

This Project demonstrates that there are many tools and procedures currently available for performing quality control in fMRI. It also presents a healthy warning that much can go wrong with the complex data acquisitions and analyses that go into fMRI, and QC should be included in all studies, whether researchers are using public datasets or acquiring their own scans. With careful preparation and quality control investigations, researchers can be more confident that their results are based on reasonable data and the intended processing. In short, we urge researchers to choose a quality control method that is thorough and understandable, and to keep looking at the data.

Author contributions

PT, DG, RR, and JE conceived of this project, organized it, and contributed to writing this article. AB and DM helped organize,

package and distribute the datasets, and further contributed in writing this article. All authors contributed to the article and approved the submitted version.

Funding

PT, DG, and RR were supported by the NIMH Intramural Research Program (ZICMH002888) of the NIH/HHS, USA. JE was supported by the National Institutes of Health, grant number: R37MH066078 to Todd S. Braver. AB and DM were supported by the NIMH Intramural Research Program (ZICMH002960) of the NIH/HHS, USA. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

Acknowledgments

We would like to thank all of the participating researchers, for spending their time and effort in processing, performing QC, and carefully presenting their results. We also thank all of the reviewers, who provided such useful feedback for the work in this Research Topic, as well as the guest editors.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

This work was written as part of PT's, DG's, RR's, and DM's official duties as US Government employees. The views expressed in this article do not necessarily represent the views of the NIH, HHS, or the United States Government.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2023.1205928/full#supplementary-material>

References

- Ardekani, B. A., and Bachman, A. H. (2009). Model-based automatic detection of the anterior and posterior commissures on MRI scans. *Neuroimage* 46, 677–682. doi: 10.1016/j.neuroimage.2009.02.030
- Ashburner, J. (2012). SPM: A history. *Neuroimage* 62, 791–800. doi: 10.1016/j.neuroimage.2011.10.025
- Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2012). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004
- Babayan, A., Erbey, M., Kumral, D., Reinelt, J. D., Reiter, A. M. F., Röbbig, J., et al. (2019). A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Sci. Data* 6, 180308. doi: 10.1038/sdata.2018.308
- Barber, A. D., Srinivasan, P., Joel, S. E., Caffo, B. S., Pekar, J. J., Mostofsky, S. H., et al. (2012). Motor "dexterity"? Evidence that left hemisphere lateralization of motor circuit connectivity is associated with better motor performance in children. *Cereb. Cortex* 22, 51–59. doi: 10.1093/cercor/bhr062
- Bernaerts, S., Prinsen, J., Dillen, C., Berra, E., Brams, S., Wenderoth, N., et al. (2016). *Oxytocin-Based Pharmacotherapy for Autism Spectrum Disorders: Investigating the Immediate and Long-Term Effects From a Neural and Behavioral Perspective*. Baltimore, MD: International Meeting for Autism Research (IMFAR).
- Bilder, R., Poldrack, R., Cannon, T., London, E., Freimer, N., Congdon, E., et al. (2018). *UCLA Consortium for Neuropsychiatric Phenomics LA5c Study*. OpenNeuro. [Dataset] doi: 10.18112/openneuro.ds000030.v1.0.0
- Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* 154, 174–187. doi: 10.1016/j.neuroimage.2017.03.020
- Clayden, J., Cox, R. W., and Jenkinson, M. (2020). *RNifti: Fast R and C++ Access to Nifti Images*. Available online at: <https://CRAN.R-project.org/package=RNifti> (accessed May 24, 2023).
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Delmonte, S., Balsters, J. H., and Gallagher, L. (2012). *Social and Monetary Reward Processing in Autism Spectrum Disorders (ASD): Interaction Effects in the Striatum*. Toronto, ON: International Meeting for Autism Research (IMFAR).
- Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 6, 659–667. doi: 10.1038/mp.2013.78
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., Gorgolewski, K. J., et al. (2017). fMRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE* 12, e0184661. doi: 10.1371/journal.pone.0184661
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIQC: A robust preprocessing pipeline for functional MRI. *Nat. Methods* 16, 111–116. doi: 10.1038/s41592-018-0235-4
- Fischl, B., and Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11050–11055. doi: 10.1073/pnas.200033797
- Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., et al. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* 54, 313–327. doi: 10.1016/j.neuroimage.2010.07.033
- Glen, D. R., Taylor, P. A., Buchsbaum, B. R., Cox, R. W., Reynolds, R. C. (2020). Beware (surprisingly common) left-right flips in your MRI data: an efficient and robust method to check MRI dataset consistency using AFNI. *Front. Neuroinformatics* 14. doi: 10.3389/fninf.2020.00018

- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* 5, 13. doi: 10.3389/fninf.2011.00013
- Goto, M., Abe, O., Aoki, S., Hayashi, N., Miyati, T., Takao, H., et al. (2013). Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra provides reduced effect of scanner for cortex volumetry with atlas-based method in healthy subjects. *Neuroradiology* 55, 869–875. doi: 10.1007/s00234-013-1193-2
- Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., et al. (2019). The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* 95, 103208. doi: 10.1016/j.jbi.2019.103208
- Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J. G., et al. (2009). Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* 42, 377–381. doi: 10.1016/j.jbi.2008.08.010
- Holla, B., Taylor, P. A., Glen, D. R., Lee, J. A., Vaidya, N., Mehta, U. M., et al. (2020). A series of five population-specific Indian brain templates and atlases spanning ages 6–60 years. *Hum. Brain Mapp.* 41, 5164–5175. doi: 10.1002/hbm.25182
- Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B., and Hoffmann, M. (2022). SynthStrip: Skull-stripping for any brain image. *Neuroimage* 260, 119474. doi: 10.1016/j.neuroimage.2022.119474
- Lee, J. S., Lee, D. S., Kim, J., Kim, Y. K., Kang, E., Kang, H., et al. (2005). Development of Korean standard brain templates. *J. Korean Med. Sci.* 20, 483–488. doi: 10.3346/jkms.2005.20.3.483
- Liu, W., Wei, D., Chen, Q., Yang, W., Meng, J., Wu, G., et al. (2017). Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. *Sci. Data* 4, 170017. doi: 10.1038/sdata.2017.17
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *Elife* 10, e71774. doi: 10.7554/eLife.71774.sa2
- McCarthy, P. (2022). *FSleyes (1.5.0)*. Zenodo. doi: 10.5281/zenodo.7038115
- Molfese, P. J., Glen, D., Mesite, L., Cox, R. W., Hoeft, F., Frost, S. J., et al. (2021). The Haskins pediatric atlas: A magnetic-resonance-imaging-based pediatric template and atlas. *Pediatr. Radiol.* 51, 628–639. doi: 10.1007/s00247-020-04875-y
- Nebel, M. B., Joel, S. E., Muschelli, J., Barber, A. D., Caffo, B. S., Pekar, J. J., et al. (2014). Disruption of functional organization within the primary motor cortex in children with autism. *Hum. Brain Mapp.* 35, 567–580. doi: 10.1002/hbm.22188
- Nieto-Castanon, A. (2020). *Handbook of Functional Connectivity Magnetic Resonance Imaging Methods in CONN*. Hilbert Press. doi: 10.56441/hilbertpress.2207.6598
- Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). *Fields: Tools for Spatial Data*. Boulder, CO: University Corporation for Atmospheric Research.
- Pekar, J. J., and Mostofsky, S. H. (2010). *FCP Classic Data Sharing Samples: Baltimore site Download*. Available online at: <http://www.nitrc.org/frs/downloadlink.php/1600> (accessed May 24, 2023).
- Petersen, S., Schlaggar, B., and Power, J. (2018). *Washington University 120. OpenNeuro*. Available online at: <https://openneuro.org/datasets/ds000243/versions/00001> (accessed May 24, 2023).
- Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K. J., Karlsgodt, K. H., Mumford, J. A., et al. (2016). A phenome-wide examination of neural and cognitive function. *Sci. Data* 3, 160110. doi: 10.1038/sdata.2016.110
- Roy, A., Bernier, R. A., Wang, J., Benson, M., French, J. J. Jr., Good, D. C., et al. (2017). The evolution of cost-efficiency in neural networks during recovery from traumatic brain injury. *PLoS ONE* 12, e0170541. doi: 10.1371/journal.pone.0170541
- Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H. M., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23(Suppl.1), S208–S219. doi: 10.1016/j.neuroimage.2004.07.051
- Snoek, L., van der Miesen, M. M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., Steven Scholte, H., et al. (2021). The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *Sci. Data* 8, 85. doi: 10.1038/s41597-021-00870-6
- Song, S., Bokkers, R. P. H., Edwardson, M. A., Brown, T., Shah, S., Cox, R. W., et al. (2017). Temporal similarity perfusion mapping: A standardized and model-free method for detecting perfusion deficits in stroke. *PLoS ONE* 12, e0185552. doi: 10.1371/journal.pone.0185552
- Tang, Y., Hojatkashani, C., Dinov, I. D., Sun, B., Fan, L., Lin, X., et al. (2010). The construction of a Chinese MRI brain atlas: A morphometric comparison study between Chinese and Caucasian cohorts. *Neuroimage* 51, 33–41. doi: 10.1016/j.neuroimage.2010.01.111
- Taylor, P., Etzel, J. A., Glen, D., Reynolds, R. C., Moraczewski, D., Basavaraj, A. (2023). *fMRI Open QC Project*. doi: 10.17605/OSF.IO/QAESM
- Wei, D., Zhuang, K., Ai, L., Chen, Q., Yang, W., Liu, W., et al. (2018). Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset. *Sci. Data* 5, 180134. doi: 10.1038/sdata.2018.134
- Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* 2, 125–141. doi: 10.1089/brain.2012.0073
- Williams, B., and Lindner, M. (2020). pyfMRIqc: A software package for raw fMRI data quality assurance. *J. Open Res. Softw.* 8, 23. doi: 10.5334/jors.280
- Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage* 92, 381–397. doi: 10.1016/j.neuroimage.2014.01.060
- Xie, Y. (2014). “KnitR: A comprehensive tool for reproducible research in R,” in *Implementing Reproducible Research, The R Series*. eds V. Stodden, F. Leisch, and R. D. Peng (Boca Raton, FL: CRC Press, Taylor & Francis Group).
- Yan, C. G., Wang, X. D., and Lu, B. (2021). DPABISurf: Data processing & analysis for brain imaging on surface. *Sci. Bulletin* 66, 2453–2455. doi: 10.1016/j.scib.2021.09.016
- Yan, C. G., Wang, X. D., Zuo, X. N., and Zang, Y. F. (2016). DPABI: Data processing and analysis for (resting-state) brain imaging. *Neuroinformatics* 14, 339–351. doi: 10.1007/s12021-016-9299-4
- Yan, C. G., and Zang, Y. F. (2010). DPARSF: A MATLAB toolbox for “pipeline” data analysis of resting-state fMRI. *Front. Syst. Neurosci.* 14, 4. doi: 10.3389/fnins.2010.00013
- Yoneyama, N., Watanabe, H., Kawabata, K., Bagarinao, E., Hara, K., Tsuboi, T., et al. (2018). Severe hyposmia and aberrant functional connectivity in cognitively normal Parkinson's disease. *PLoS One* 13, e0190072. doi: 10.1371/journal.pone.0190072



OPEN ACCESS

EDITED BY

Daniel R. Glen,
National Institute of Mental Health
(NIH), United States

REVIEWED BY

Andrew Jahn,
University of Michigan, United States
Jonathan Ipser,
University of Cape Town, South Africa

*CORRESPONDENCE

Xin Di

✉ xin.di@njit.edu

Bharat B. Biswal

✉ bbiswal@yahoo.com

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroimaging

RECEIVED 14 October 2022

ACCEPTED 19 December 2022

PUBLISHED 10 January 2023

CITATION

Di X and Biswal BB (2023) A functional
MRI pre-processing and quality
control protocol based on statistical
parametric mapping (SPM) and
MATLAB.

Front. Neuroimaging 1:1070151.
doi: 10.3389/fnimg.2022.1070151

COPYRIGHT

© 2023 Di and Biswal. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

A functional MRI pre-processing and quality control protocol based on statistical parametric mapping (SPM) and MATLAB

Xin Di* and Bharat B. Biswal*

Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ,
United States

Functional MRI (fMRI) has become a popular technique to study brain functions and their alterations in psychiatric and neurological conditions. The sample sizes for fMRI studies have been increasing steadily, and growing studies are sourced from open-access brain imaging repositories. Quality control becomes critical to ensure successful data processing and valid statistical results. Here, we outline a simple protocol for fMRI data pre-processing and quality control based on statistical parametric mapping (SPM) and MATLAB. The focus of this protocol is not only to identify and remove data with artifacts and anomalies, but also to ensure the processing has been performed properly. We apply this protocol to the data from fMRI Open quality control (QC) Project, and illustrate how each quality control step can help to identify potential issues. We also show that simple steps such as skull stripping can improve coregistration between the functional and anatomical images.

KEYWORDS

functional MRI, head motion, pre-processing, quality control, resting-state, skull stripping

1. Background

Functional MRI (fMRI), especially blood-oxygen-level dependent (BOLD) fMRI (Ogawa et al., 1992), has become a popular technique to study brain functions underlying cognitive and affective processes, and to investigate brain alterations in psychiatric and neurological disorders. The sample sizes of fMRI studies have been steadily increasing over the years (Poldrack et al., 2017; Yeung et al., 2020), and many researchers have taken advantages of large open-access datasets, such as 1,000 Functional Connectomes Project (Biswal et al., 2010), autism brain imaging data exchange (ABIDE) (Di Martino et al., 2014), Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2015), and OpenNeuro (Markiewicz et al., 2021). The wide availability and the heterogeneity in acquisition protocols and data quality make it challenging for data processing and statistical analysis. Quality control on the data processing has become a critical component in research but has not been fully charted.

The quality assurance for an fMRI study span from data acquisition to data processing and statistical analysis (See [Lu et al., 2019](#) for an overview). Here we assume that the data have already been collected or obtained from an online repository. Then the quality assurance starts with checking the quality of the images, and mainly involves the data processing steps. There are automated quality control measures for specific steps, e.g., assessing the quality of MRI images ([Esteban et al., 2017](#)) and brain registration ([Benhajali et al., 2020](#)). But published studies on quality control usually do not cover the entire processing pipeline. In this paper, we outline a processing pipeline for fMRI data that has been used in our lab, and detail the quality control procedure after each of the pre-processing steps. The pre-processing pipeline is suitable for all resting-state, task state, and movie watching conditions ([Di and Biswal, 2019, 2020, 2022; Di et al., 2020, 2022a,b](#)). The protocol is based on Statistical Parametric Mapping (SPM) (<https://www.fil.ion.ucl.ac.uk/spm/>) under MATLAB environment. The quality control issues may be similar when using other major software, e.g., Analysis of Functional NeuroImages (AFNI) ([Cox, 1996](#)) and the FMRIB Software Library (FSL) ([Jenkinson et al., 2012](#)). But the implementations of quality control in other software are outside the scopes of this paper.

Quality control is mainly 2 fold. The first is to identify artifacts and issues in the images. This includes spatial domain issues, such as ghost artifacts, lesions, and brain coverage, as well as temporal domain issues, such as head motion and other physiological noises. The second is to ensure that the data processing steps can run properly. Practically, many data processing steps rely on iterations, which are sensitive to initial conditions. Quality control is critical to ensure that these processing steps can run properly but are not stuck in local minima. In addition, given the complexity of the fMRI data, there might always be unexpected issues in the images or different processing steps. Visualizations of different aspects of the images will always be helpful to spot the unexpected issues.

Here, we first describe the pre-processing and quality control protocol in detail, including visualizations, exclusion criteria, and the steps needed for processing assurance. The protocol mainly relies on SPM and MATLAB functions. Some visualizations are inspired by previous works, such as TSDiffana (<http://imaging.mrc-cbu.cam.ac.uk/imaging/DataDiagnostics>) and [Power et al. \(2014\)](#). And secondly, we apply the protocol to the data of the Open QC Project (<https://osf.io/qaesmf/>). We illustrate how quality issues can be identified, and what steps are needed to ensure proper data processing. One particular step is the usage of skull-stripped anatomical images for functional-anatomical images coregistration ([Fischmeister et al., 2013](#)). By using the OpenQC dataset, we examine how skull stripping can potentially improve the coregistration compared with using the raw anatomical images.

2. Pre-processing and quality control protocol

2.1. Software

SPM12: v7771 under MATLAB R2021a environment.

2.2. Procedure

The outline of the pre-processing and quality control steps is shown in [Figure 1](#). The codes are available at https://github.com/Brain-Connectivity-Lab/Preprocessing_and_QC.

2.2.1. Q1. Data initial check

The purposes of the initial check include checking the consistency of imaging parameters across participants, and checking the image quality, coverage, and orientations of the functional and anatomical images.

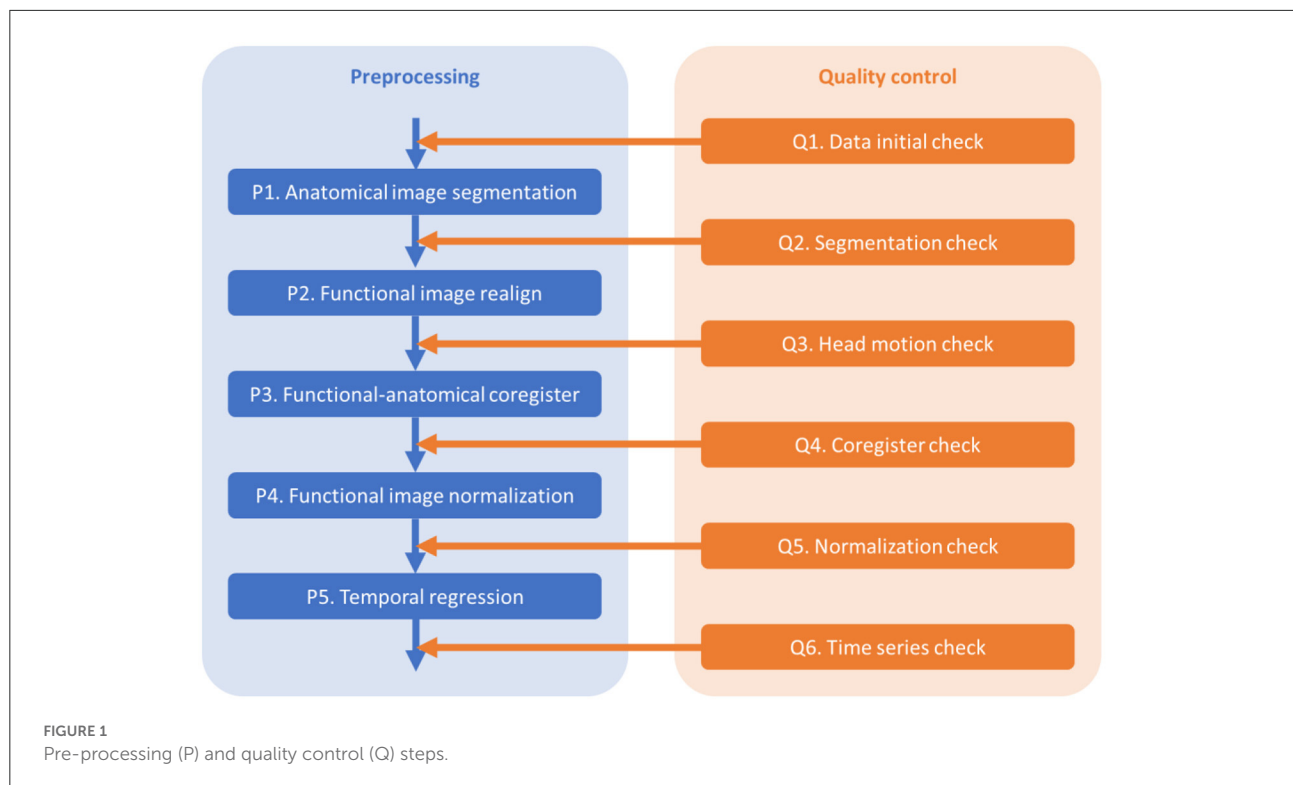
First, check the key parameters that may affect pre-processing, including the number of volumes, repetition time (TR), and voxel sizes. Plot the parameters across participants (e.g., [Supplementary Figure 1](#)) or the histograms may be helpful. If a few participants have different parameters, e.g., fewer volumes, they may be removed from further analyses. If many participants have various numbers of volumes, one may consider keeping the same number of volumes across all the participants. Otherwise, one may also consider including covariates in group level models to account for the parameter variations.

Second, check the anatomical images using SPM Check Registration functionality. The first image is the anatomical image of a participant in native space, and the second is the single subject T1 weighted template image in MNI space ([Figure 2A](#)). The contour of the first image can be overlayed onto the second image. Focus on, (1) whether the anatomical image has the same orientation and similar initial position to the template, (2) any artifacts, e.g., ghosting, and brain lesion. If any anomaly is noted, then the image needs to be further checked for the whole brain volume. If the anatomical image is located far from the MNI template, or rotated into a different orientation, then manually reorient the image to the template direction and reset the origin to the anterior commissure.

Thirdly, check the first functional image using SPM Check Registration functionality. This is the same as the previous step, except that the first image is a functional image. Focus on (1) whether the functional image has the same orientation and similar initial position to the MNI space template, (2) any artifacts, e.g., ghosting, and (3) the spatial coverage.

2.2.2. P1. Anatomical image segmentation

The purpose of this step is to segment the anatomical image of a participant into gray matter (GM), white matter (WM),



cerebrospinal fluid (CSF), and other tissues, and obtain the parameters (deformation fields) for the spatial normalizations of the functional images. A bias corrected anatomical image is also generated, which will be used for functional-anatomical image registration.

Use SPM Segment functionality. The input volume is the subject's anatomical image. Additional non-default setting: (1) "Save Bias Corrected" -> "Save Bias Corrected"; (2) "Warped Tissue" for the first three tissue types (GM, WM, and CSF) -> "Unmodulated"; and 3) "Deformation Fields" -> "Forward."

DARTEL (a fast diffeomorphic registration algorithm) may be used to generate a sample specific template for spatial normalization (Ashburner, 2007). It can improve cross-individual registrations, especially for a homogeneous sample from a specific population, e.g., children or old adults. But for a large sample size with diverse demographics, DARTEL may not be necessary and is computationally expensive.

2.2.3. Q2. Anatomical image segmentation check

The purpose of this step is to check the quality of segmentation.

Use SPM Check Registration functionality. The first image is the segmented gray matter density image in MNI space (wc1xxx), and the second image is the single subject T1 weighted image in MNI space (Figure 2B). The contour of the first image

can be overlaid onto the second image. Next, overlay the segmented images of GM, WM, and CSF (wc1xxx, wc2xxx, and wc3xxx) to the first image.

If misclassification of any tissues is noted, then double check the original anatomical image. If the misclassification could be caused by the position/orientation of the raw anatomical image, try to manually reorient the anatomical image. If brain lesions or image quality issues are noticed, this participant's data should be excluded.

2.2.4. P2. Functional images realign

The purpose of this step is to align all the functional images of a run to the first image. Rigid body head motion parameters (rp files) are also obtained.

Use SPM Realign: Estimate & Reslice functionality. For "Data:Session": input all the functional images. Non-default setting: "Resliced images": "Mean Image Only."

2.2.5. Q3. Head motion check

The purpose of this step is to check the distributions of head motion in the sample, and remove participants with excessive head motions from further analyses.

Calculate framewise displacement (FD) in translation and rotation based on the rigid body transformation results from the

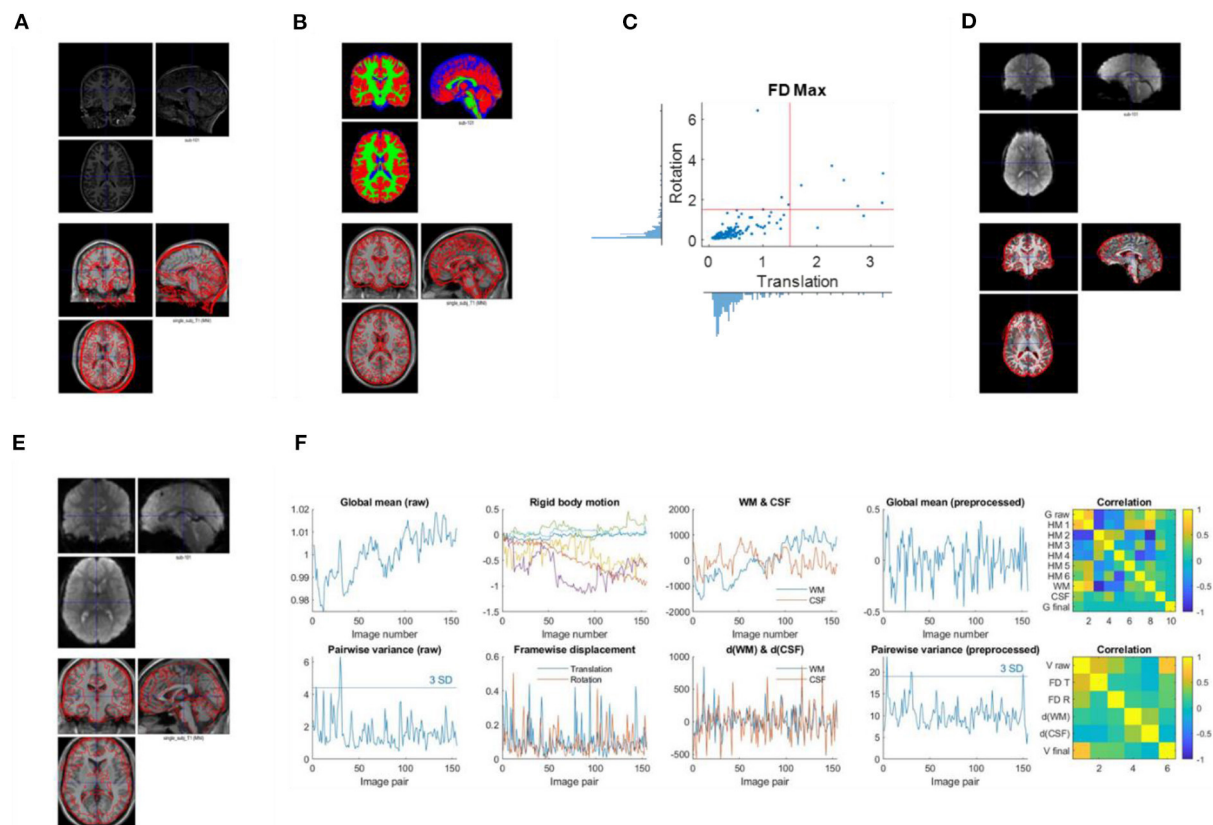


FIGURE 2

Example visualizations of each quality control step. (A) Image initial check (Q1). (B) Segmentation check (Q2). (C) Head motion check (Q3). (D) Coregister check (Q4). (E) Normalization check (Q5). (F) Time series check (Q6).

P2 step (Di and Biswal, 2015). The formula for FD at time t are as follows,

$$FD_{translation,t} = \sqrt{(hp_{x,t} - hp_{x,t-1})^2 + (hp_{y,t} - hp_{y,t-1})^2 + (hp_{z,t} - hp_{z,t-1})^2}$$

$$FD_{rotation,t} = \sqrt{(hp_{\alpha,t} - hp_{\alpha,t-1})^2 + (hp_{\beta,t} - hp_{\beta,t-1})^2 + (hp_{\gamma,t} - hp_{\gamma,t-1})^2}$$

Where hp represents the head position parameters estimated relative to the first image. x , y , and z represent the three translation directions, and α , β , and γ represent the three rotation directions. Plot the distributions of maximum framewise displacement across all the participants (Figure 2C). A pre-specified threshold of maximum framewise displacements >1.5 mm or 1.5° (approximately half of the voxel sizes) can be used to exclude participants. However, the threshold may depend on the sample characteristics. See below for more discussions.

2.2.6. P3. Functional-anatomical images coregister

The purpose of this step is to coregister the functional images to the anatomical image of the same individual.

First, generate a skull-stripped bias-corrected anatomical image using SPM Image Calculator (ImCalc) functionality. Input Images: (1) the bias-corrected anatomical image, (2) through (4) c1xxx, c2xxx, and c3xxx segmented tissue images, respectively. Expression: $i1.*((i2+i3+i4)>0.5)$.

Second, use SPM Coregister:Estimate functionality. “Reference Image”: the skull-stripped bias-corrected anatomical image; “Source Image”: the mean functional image generated in the realign step; “Other Images,” all the functional images of the run.

2.2.7. Q4. Coregistration check

The goal of this step is to check the quality of coregistration between the functional and anatomical images.

Use SPM Check Registration functionality. The first image is a functional image of a participant in native space, and second

image is the skull stripped anatomical image in native space (Figure 2D). The contour of the first image can be overlayed onto the second image.

Check whether the contour of the functional image aligns with the anatomical image. If the two images are not aligned well, then manual reorientation of the images may be needed.

2.2.8. P4. Spatial normalization

The purpose of this step is to spatially normalize all the functional images into the common MNI space. The normalization parameters are obtained from the segmentation step.

Use SPM Normalize/Write functionality. “Data:Subject:Deformation Field”: y_xxx file from the anatomical image folder; “Images to Write”: all the functional images of a run. Non-default setting, “Voxel sizes”: $3 \times 3 \times 3$. The resampling voxel size should be similar to the original voxel size. For the fMRI QC data, we used a common voxel size of $3 \times 3 \times 3 \text{ mm}^3$. This may be modified according to the actual voxel size. The resampled voxel size also affects the estimates of spatial smoothness, which may in turn affect voxel-wise statistical results (Mueller et al., 2017).

2.2.9. Q5. Normalization check

The purpose of this step is to check the spatial registrations of the fMRI images to an MNI space template.

Use SPM Check Registration functionality. The first image is the normalized functional image of a participant in MNI space, and the second image is the single subject anatomical image in MNI space. The contour of the first image can be overlayed onto the second image (Figure 2E).

2.2.10. P5. Voxel-wise general linear model

For resting-state data, this step is used to regress out variations of no-interest, such as low-frequency drift, head motion, and WM/CSF signals. The residual images will be further used to calculate functional connectivity or resting-state parameters, such as amplitude of low-frequency fluctuations (ALFF) (Yang et al., 2007), regional homogeneity (ReHo) (Zang et al., 2004), and physiophysiological interaction (PPI) (Di and Biswal, 2013). For task fMRI, the purpose of this step is mainly to derive task related activations.

For resting-state data, firstly, define WM and CSF masks by thresholding and resampling the subject's segmented tissue images using SPM Image Calculator (ImCalc) functionality. “Input Images”: (1) the first functional image (to define the voxel dimension), and (2) $wc2xxx$ or $wc3xxx$ normalized tissue density image. “Expression”: $i2 > 0.99$. The threshold is used to ensure only WM or CSF voxels are included in the masks.

Secondly, extract the first principal component of the signals in the WM and CSF masks, respectively.

Thirdly, use General Linear Model (GLM) functionality in SPM to perform the regression. The regressors include 24 Friston's head motion model (Friston et al., 1996), the first PC of the WM and CSF, respectively, and a constant term. Note that an implicit high pass filter is also included in the GLM with a cut-off of 1/128 Hz. This GLM step essentially performs artifact removal and filtering together, which can prevent introduced artifacts when doing these two steps separately (Lindquist et al., 2019).

Fourthly, estimate the GLM using SPM Model estimation functionality. Non-default setting, “Write residuals”: Yes.

For task fMRI data, also use the GLM functionality in SPM to perform the regression. Define task regressors using the design timing parameters. Additional regressors include 24 Friston's head motion model (Friston et al., 1996) and a constant term. Note that an implicit high pass filter is also included in the GLM with a cut-off of 1/128 Hz. Next, estimate the GLM using SPM Model estimation functionality. The residual images can be saved to check model fitness, but usually they are not needed for further analyses.

2.2.11. Q6. Time series check

For resting-state data, the purpose of this step is to check the time series of global signals, and their relations to head motion and physiological noises. Mean global signals and pairwise variance [similar to DVARS, temporal derivative of variance (Power et al., 2014)] are commonly used to quality control fMRI time series. Outliers of the variance time series are usually caused by head motion. Therefore, plotting head motion parameters together with the variance and global signals can help to illustrate the relationships. A further question is whether the linear regression step can effectively minimize the noises in the global signals.

Plot time series as Figure 2F. Top row, first, the global mean intensity for the raw fMRI images; second, six rigid body head motion parameters in mm or degree; third, the first PC of the signals in the WM and CSF; and fourth, the global mean intensity for the pre-processed fMRI images within a brain mask. The correlations among all these time series are shown in the last column. Bottom row, first, pairwise variance between consecutive images for the raw fMRI images; second, framewise displacement in translation and rotation; third, derivative (difference) of the first PCs in WM and CSF; and fourth, pairwise variance between consecutive images from the pre-processed fMRI images within a brain mask. The correlations among all these time series are shown in the last column.

The pairwise variance time series is a simple way to spot extreme values. One can use three standard deviations as a criterion to identify the extreme values. Similar spikes can usually be seen in the framewise displacement time series, and

TABLE 1 Key imaging parameters in the eight sites of the fMRI Open QC project.

Site	<i>n</i>	Number of functional volumes	TR (s)	Functional image voxel size			Anatomical image voxel size		
				x	y	z	x	y	z
Task	30	242	2	3	3	4	1	1	1
Rest 1	20	128 or 156	2.5	2.67 or 2.29	2.67 or 2.29	3	1	1	1
Rest 2	20	150	2	3	3	3.84	1	1 or 0.93	1 or 0.93
Rest 3	16	162	2.5	1.56	1.56	3.1	0.98	1.2	0.98
Rest 4	23	123	2.5	2.67	2.67	3	1	1	1
Rest 5	20	144	2	3 or 1.85	3 or 1.85	4	1	1	1
Rest 6	20	130–724	2.5	4	4	4	1	1	1
Rest 7	20	198	2.5	3	3	3.51	1	1	1

Shaded cells indicate the presence of different parameters within the site. TR, repetition time.

sometimes are also visible in the derivatives of the WM/CSF signals. This will result in high correlations among the pairwise variance, framewise displacement, and WM/CSF derivatives. Also focus on the pairwise variance time series from the pre-processed images to check whether they are no longer correlated with the framewise displacement or WM/CSF derivatives. A threshold, e.g., $r > 0.3$, can be used to identify large correlations.

For task data, the effects of interest are usually the brain activity related to the task design. Then the focus of this step is to check whether the global signals and head motions are correlated with the task design. Therefore, in addition to the time series of global signals and head motion, also plot the task design time series and their derivatives. If the global signals or head motion parameters are correlated with the task design, or the pair wise variance or framewise displacement are correlated with the derivatives of the task design, then one may consider controlling these factors in the first level GLMs.

2.3. Other processing steps

Spatial smoothing is not included in this protocol. It is only necessary when voxel-wise statistical analysis is used. If the analysis is ROI based connectivity analysis, then smoothing is not necessary. Moreover, when calculating ReHo, which is a commonly used resting-state measure, the data should also be un-smoothed.

3. Materials and methods

3.1. Datasets

The data were obtained from the fMRI Open QC Project (<https://osf.io/qaesm/>). There are anatomical and functional MRI data of 169 participants from eight sites. Seven sites are resting-state fMRI, and the remaining one is task-based

TABLE 2 FMRI quality control criteria.

FMRI quality control criteria	Exclude a subject if:
A. Imaging parameters	Deviating from other participants
B. Anatomical image quality and coverage	Visual assessment
C. Functional image quality and coverage	Visual assessment
D. Segmentation failure	Visual assessment
E. Maximum framewise displacement	>Than 1.5 mm or 1.5°

fMRI. The data were aggregated from different online resources, including 1,000 Functional Connectomes Project (Biswal et al., 2010), ABIDE (Di Martino et al., 2014), and OpenNeuro (Markiewicz et al., 2021).

The MRI images were acquired using different MRI scanners and imaging protocols. All the MRI scanners were 3T. Table 1 lists some key parameters useful for data analyses. Note that some parameters vary within a site. More imaging parameters for all the participants are shown in Supplementary Figure 1.

3.2. Pre-processing and quality control

We followed the protocol outlined in Section 2. For each quality control step, an image was saved for each subject. The output images were visually inspected across all the participants. The quality control and exclusion criteria are summarized in Table 2.

3.3. Data analysis

In the functional-anatomical images coregister step, the current protocol uses the bias-corrected skull-stripped anatomical image as the reference. Because the signals

in the skull in the EPI images are weak, in theory it is preferable to coregister the functional images to the skull-stripped anatomical images. However, this is not the default recommendation in SPM. A study has suggested that using the skull-stripped image may improve group-level statistical results (Fischmeister et al., 2013). However, no formal comparison has been performed. We hypothesize that in most cases using the non-skull stripped images will perform the same as the skull stripped images. However, in a small number of cases, using the raw anatomical image may fail. By using the fMRI QC dataset, we estimate the number of cases that would fail when using the raw anatomical image as the reference.

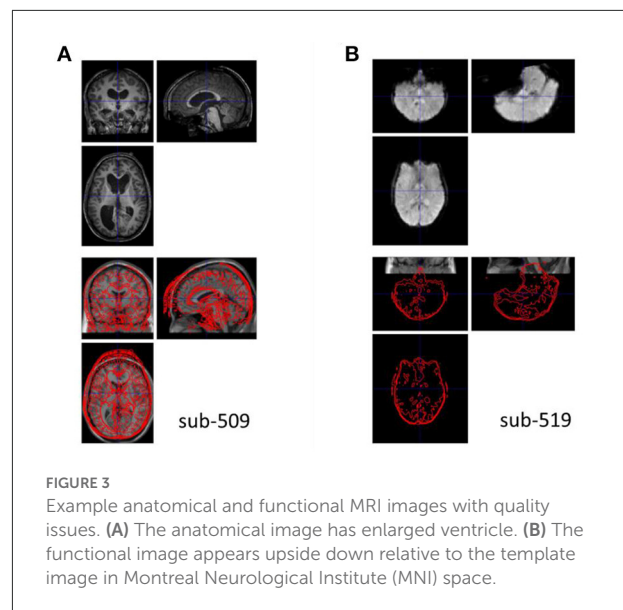
Specifically, we also performed the coregister step by using the raw anatomical image as the reference. We calculated the spatial distance between the functional images to the different reference images. The Euclidean distances were calculated in translation and rotation, separately. We used a threshold of 9 mm or 9° (~ 3 voxels) to identify cases with excessive differences. We then overlaid the two functional images with the anatomical images to identify potential causes of the discrepancy.

4. Results

4.1. Q1. Data initial check

Supplementary Table 1 shows some key imaging parameters of the functional and anatomical images for every participant. In resting-state site 1, two participants had fewer fMRI volumes than the rest of the group, which should be removed from analysis. In resting-state site 6, the numbers of fMRI volumes varied between 130 and 724. We kept the first 130 volumes from all the participants for further analysis. The voxel sizes of fMRI images in resting-state site 1 and site 5 varied across participants. Given that only a few participants had different voxel sizes from the majority participants of a site, these participants should be removed from further analysis. The voxel sizes of the anatomical images in resting-state site 2 also varied across the participants. However, it may have minimum impact on the functional images and were therefore kept for further analysis.

The anatomical images were visually inspected for their quality, coverage, and relative positions to the MNI template. All the images were close to the MNI template, indicating that no manual origin setting was needed. One participant's image (sub-509) showed enlarged ventricles (Figure 3A), which should be removed from further analysis. Another participant's image (sub-203, not shown) had mildly enlarged ventricle, which extended to the right lingual territory. We classified this participant as uncertain. This participant may



be included if the visual areas were not the main regions of interest.

The quality and coverage of the first fMRI images seemed acceptable for all the participants. However, two participants' images (sub-518, sub-519) appeared upside down (e.g., Figure 3B). The images were manually reoriented to the template orientation.

4.2. Q2. Anatomical image segmentation check

The segmentation procedure seriously failed in two participants (sub-509 and sub-511). For sub-509, most gray matter regions were identified as CSF (Figure 4A). And for sub-511, part of the visual gray matter was missing, and no CSF was identified (Figure 4B).

Five other participants (sub-108, sub-405, sub-420, sub-512, and sub-514) also have minor segmentation issues, particularly in the CSF (e.g., Figure 4C). Since fMRI analysis usually focuses on gray matter, the misclassifications of CSF may not affect the normalizations of gray matter. These participants may be included in the following analysis. We labeled them uncertain because they may not be included in other types of analysis, such as voxel-based morphometry (Ashburner and Friston, 2000).

4.3. Q3. Head motion and variance check

When using the pre-specified threshold of maximum framewise displacement > 1.5 mm or 1.5° , 12 participants were removed from further analysis. Figure 5 shows the distributions

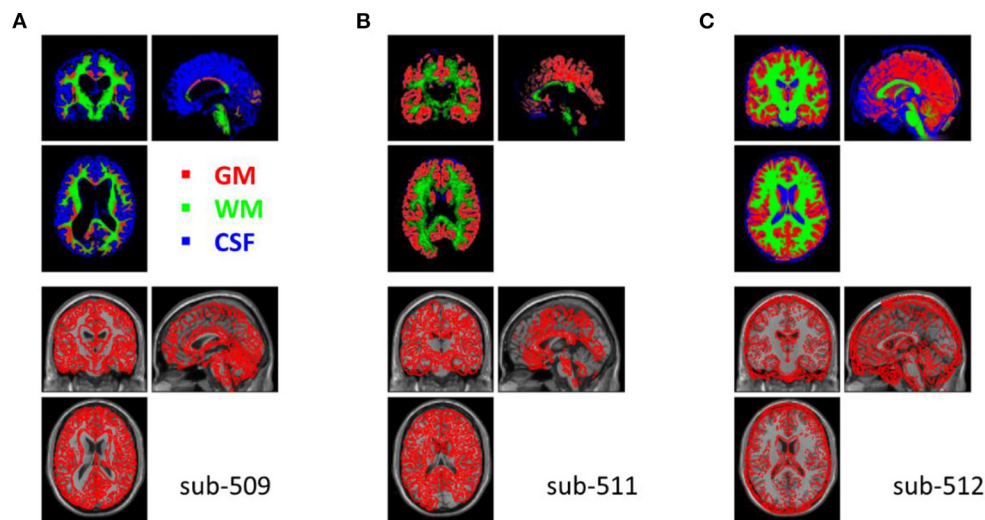


FIGURE 4

Example anatomical images with segmentation issues. Top row shows segmented tissue images of gray matter (red), white matter (green), and cerebrospinal fluid (blue) in Montreal Neurological Institute (MNI) space. Bottom row shows the single subject T1 image in MNI space with the segmented gray matter contours. (A) Shows the participant where most of the gray matter was misclassified as CSF. (B) Shows missing classified gray matter in the visual cortex and no classifications of CSF. (C) Shows that many soft tissues and bones outside the cortex were misclassified as CSF.

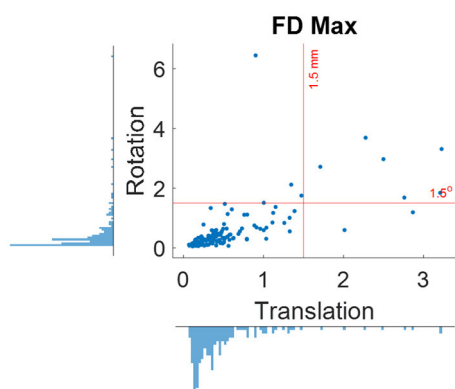


FIGURE 5

Distributions of maximum framewise displacement (FD) in translation and rotation. The red lines indicate the 1.5 mm and degree thresholds used for excluding participants.

of maximum framewise displacement across all the participants. It appears that the 1.5 mm and 1.5° threshold only remove a few participants with excessive head motions. This is desirable because the removal is supposed to only apply to outliers.

4.4. Q4. Functional-anatomical images coregister

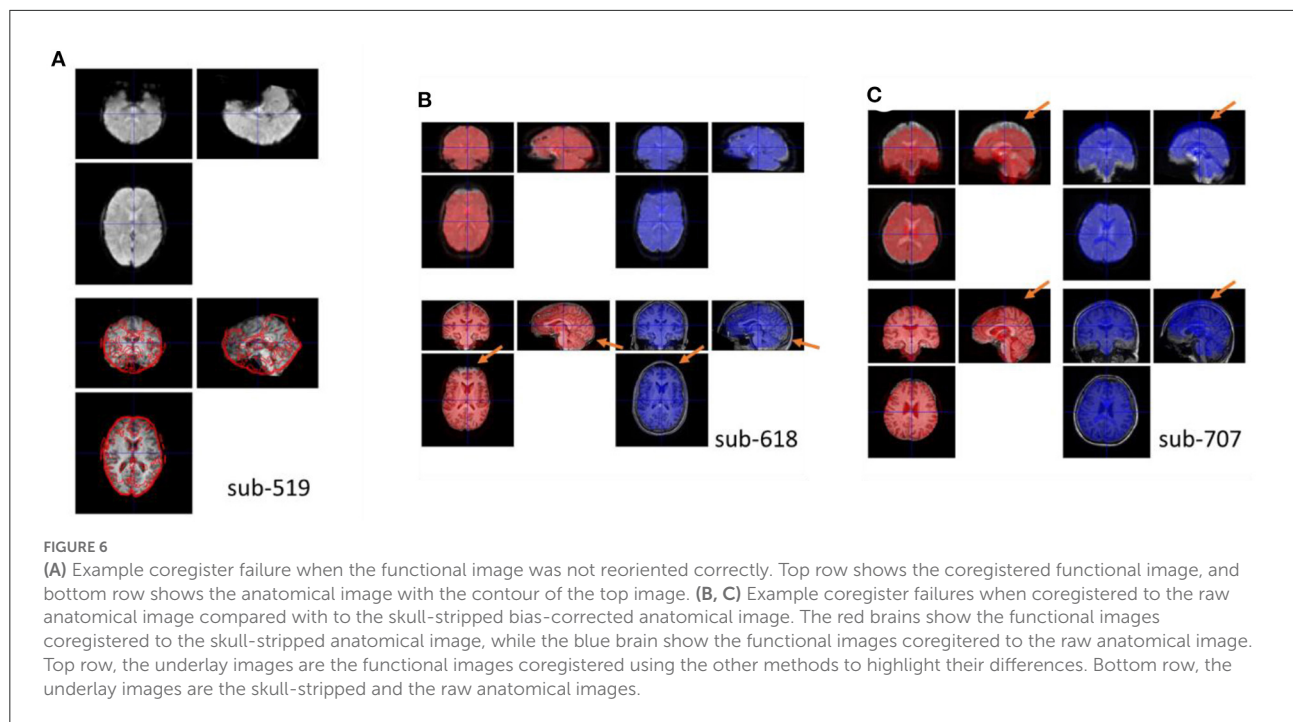
For all the participants, the functional images were properly coregistered to their respective anatomical images.

This was achieved with the previous quality assurance steps. For example, if the upside-down functional images (sub-518 and sub-519) were not manually reoriented, the coregistration step would fail. Figure 6A shows an example of a functional image registered upside-down with the anatomical image, which was stuck at a local minimum.

Moreover, if the raw anatomical image was used as a reference, the functional images may mis-aligned with the anatomical image in many participants. Figures 6B, C shows two examples. In Figure 6C, the top edge of the fMRI image was aligned to the skull when registered to the raw anatomical image. This is a typical scenario of misalignment. In Figure 6B, the functional image has a signal dropout in the prefrontal region. The distorted prefrontal edge was aligned with un-distorted prefrontal edge in the anatomical image, which resulted in a misalignment. This can be prevented by using the skull-stripped image as the reference. For each participant, we calculated spatial distance in translation and rotation between the functional images coregistered using the two reference images (Supplementary Figure 2). Four participants (2.4%) had spatial distance larger than 9 mm.

4.5. Q6. Normalization

All participants' data were successfully normalized into the MNI space.



4.6. Q7. Time series check

Figure 7 shows an example participant with large head motions. Both the global mean signals (Figure 7A) and pairwise variance (Figure 7F) showed a spike at around the 50th image. The rigid body motion parameters (Figure 7B) and framewise displacement (Figure 7G) showed similar spikes. However, the shapes of the spikes in the rigid body motion parameters appeared different from the global signals (Figure 7A), indicating that simply regressing out the rigid body parameters cannot fully remove motion related noises. In contrast, framewise displacement (Figure 7G) showed strikingly similar patterns as the pairwise variance (Figure 7F). Similarly, the rigid body movement related changes can be seen in the WM signals (Figure 7C), but only the derivatives (Figure 7H) showed similar spike patterns as the pairwise variance (Figure 7F). Next, we check whether the GLM step has successfully minimized the motion related components in the fMRI signals. The global mean signals of the pre-processed images (Figure 7D) no longer contained the spike, and so did the pairwise variance time series (Figure 7I). This is supported by the fact that the pairwise variance from the pre-processed data was not correlated with framewise displacement, which contrasted with the pairwise variance from the raw data (Figure 7J). This suggests that the GLM process can effectively minimize head motion effects in this participant, even though this participant was excluded with our pre-specified threshold.

Figure 8 shows an example participant with large head motions from the task data. The head motion effects were not

clearly present in the global mean signals (Figure 8A), but can be clearly seen in the pairwise variance time series (Figure 8E), which can be confirmed in the rigid body motion parameters (Figure 8B) and framewise displacement time series (Figure 8F). For the task-based fMRI, it is critical to verify whether head motion is related to the task design. In Figures 8C, G, we plotted the time series of task design and their derivatives. It seems that head motions were not correlated with the task design, which can be further confirmed in Figures 8D, H.

4.7. Summary of quality control results

In total, two participants were discarded due to missing time points; five were discarded due to different fMRI voxel sizes; one was discarded due to poor anatomical image quality; one was discarded due to segmentation failure; and 11 were discarded due to large head motions. Another 5 participants' data had mild issues in the anatomical images or tissue segmentations, which were marked as uncertain. A list of all the excluded or uncertain participants and their reasons is summarized in Supplementary Table 1.

5. Discussion

In this paper, we outlined a protocol for fMRI pre-processing and quality control based on SPM and MATLAB. We applied the protocol to the fMRI Open QC dataset, and identified

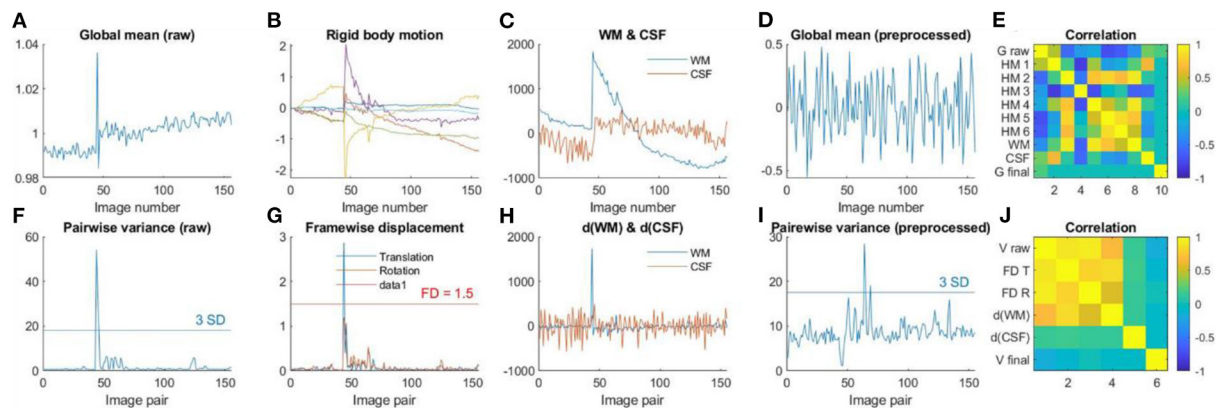


FIGURE 7

(A) Global mean intensity for the raw fMRI images. (B) Six rigid-body head motion parameters in mm or degree. (C) The first principal component (PC) of the signals in the white matter (WM) and cerebrospinal fluid (CSF). (D) Global mean intensity for the pre-processed fMRI images within a brain mask. (E) Correlations among (A) through (D). (F) Variance between consecutive images from the raw data. (G) Framewise displacement (FD) in translation and rotation. (H) Derivatives of the first PCs in WM and CSF. (I) Variance between consecutive images from the pre-processed fMRI images within a brain mask. (J) Correlations among (F) through (I).

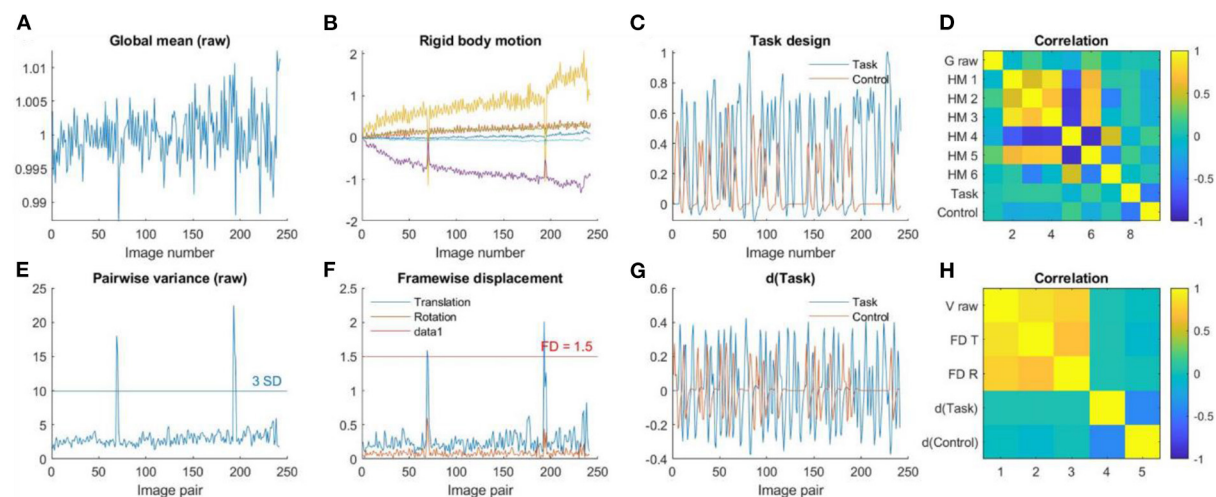


FIGURE 8

(A) Global mean intensity for the raw fMRI images. (B) Six rigid-body head motion parameters in mm or degree. (C) The task design regressors of the Task and Control conditions. (D) Correlations among (A) through (C). (E) Variance between consecutive images from the raw data. (F) Framewise displacement in translation and rotation. (G) Derivatives of the task design regressors. (H) Correlations among (E) through (G).

quality issues after each step of pre-processing. We also demonstrated that quality control can ensure proper processing. And specifically, using the skull-stripped anatomical image can help to effectively prevent mis-registrations between functional and anatomical images.

Using a skull-stripped anatomical image as a reference in the coregister step is not the default setting in SPM, but the SPM manual does recommend that if the step is unsuccessful then the skull-stripped images should be used. The current analysis showed that only a small portion of participants have failed this step. However, because they are rare, they are easily overlooked. And in some cases, e.g., Figure 6B, it is not easy to spot the

failure visually unless the two functional images are overlaid directly over each other. On the other hand, making the skull-stripped image only takes one simple step with minimal time and computation efforts. Therefore, we recommended that the skull strip should always be performed.

Head motion is a major factor that affect fMRI signals (Friston et al., 1996) and functional connectivity measures (Power et al., 2012; Van Dijk et al., 2012). Different methods have been developed to detect and minimize head motion related artifacts (Friston et al., 1996; Muschelli et al., 2014; Power et al., 2014, 2019). The Friston's 24 model has been shown to be an effective way to reduce motion related artifacts

(Yan et al., 2013), which is adopted in the current protocol. In addition to correcting motion related artifacts from the fMRI data, identifying and excluding participants with excessive head motion are also critical. In the current protocol, we set a threshold of 1.5 mm and 1.5° to remove participants with excessive head motions. We note that the threshold is arbitrary. More critically, the distributions of head motion in a sample should always be checked. If the overall head motions are large in the sample, then a more lenient threshold may be considered. If there are multiple groups, e.g., case and control, the distributions of head motion should be compared between groups. Any group differences may need to be controlled in the group-level statistical models. But one needs to keep in mind that excluding participants with large head motion may introduce sampling bias (Kong et al., 2014; Nebel et al., 2022).

Lastly, we note that the quality and formats of fMRI data varied greatly from different sources. We have only demonstrated a handful of quality issues that are present in the fMRI QC project. There are always unexpected issues when processing new data, especially when data are derived from online repositories. Making visualizations of different aspects of the data (e.g., images and time series) is always helpful to ensure proper data processing and to spot unexpected issues.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://osf.io/qaesm/>.

Author contributions

XD performed the analysis and wrote the first draft of the manuscript. XD and BB contributed to conception of the study and manuscript revision, read, and approved the submitted version.

References

- Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007
- Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582
- Benhajali, Y., Badhwar, A., Spiers, H., Urchs, S., Armoza, J., Ong, T., et al. (2020). A Standardized protocol for efficient and reliable quality control of brain registration in functional MRI studies. *Front Neuroinform.* 14, 7. doi: 10.3389/fninf.2020.00007
- Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res. Int. J.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Di, X., and Biswal, B. B. (2013). Modulatory interactions of resting-state brain functional connectivity. *PLoS ONE* 8, e71163. doi: 10.1371/journal.pone.0071163
- Di, X., and Biswal, B. B. (2015). Characterizations of resting-state modulatory interactions in the human brain. *J. Neurophysiol.* 114, 2785–2796. doi: 10.1152/jn.00893.2014
- Di, X., and Biswal, B. B. (2019). Toward task connectomics: examining whole-brain task modulated connectivity in different task domains. *Cereb. Cortex* 29, 1572–1583. doi: 10.1093/cercor/bhy055

Funding

This study was supported by (US) National Institute of Mental Health grants to XD (R15MH125332) and BB (R01MH131335).

Acknowledgments

The authors would like to thank Donna Chen for her comments on earlier versions of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnimg.2022.1070151/full#supplementary-material>

- Di, X., and Biswal, B. B. (2020). Intersubject consistent dynamic connectivity during natural vision revealed by functional MRI. *Neuroimage* 2020, 116698. doi: 10.1016/j.neuroimage.2020.116698
- Di, X., and Biswal, B. B. (2022). Principal component analysis reveals multiple consistent responses to naturalistic stimuli in children and adults. *Hum. Brain Mapp.* 43, 3332–3345. doi: 10.1002/hbm.25568
- Di, X., Woelfer, M., Kühn, S., Zhang, Z., and Biswal, B. B. (2022a). Estimations of the weather effects on brain functions using functional MRI: a cautionary note. *Hum. Brain Mapp.* 43, 3346–3356. doi: 10.1002/hbm.25576
- Di, X., Zhang, H., and Biswal, B. B. (2020). Anterior cingulate cortex differently modulates frontoparietal functional connectivity between resting-state and working memory tasks. *Hum. Brain Mapp.* 41, 1797–1805. doi: 10.1002/hbm.24912
- Di, X., Zhang, Z., Xu, T., and Biswal, B. B. (2022b). Dynamic and stationary brain connectivity during movie watching as revealed by functional MRI. *Brain Struct. Funct.* 227, 2299–2312. doi: 10.1101/2021.09.14.460293
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE* 12, e0184661. doi: 10.1371/journal.pone.0184661
- Fischmeister, F., Ph., S., Höllinger, I., Klinger, N., Geissler, A., Wurnig, M. C., et al. (2013). The benefits of skull stripping in the normalization of clinical fMRI data. *Neuroimage Clin.* 3, 369–380. doi: 10.1016/j.nicl.2013.09.007
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312
- Jack, C. R., Barnes, J., Bernstein, M. A., Borowski, B. J., Brewer, J., Clegg, S., et al. (2015). Magnetic resonance imaging in Alzheimer's disease neuroimaging initiative 2. *Alzheimers Dement.* 11, 740–756. doi: 10.1016/j.jalz.2015.05.002
- Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). *FSL NeuroImage*. 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Kong, X., Zhen, Z., Li, X., Lu, H., Wang, R., Liu, L., et al. (2014). Individual differences in impulsivity predict head motion during magnetic resonance imaging. *PLoS ONE* 9, e104989. doi: 10.1371/journal.pone.0104989
- Lindquist, M. A., Geuter, S., Wager, T. D., and Caffo, B. S. (2019). Modular pre-processing pipelines can reintroduce artifacts into fMRI data. *Hum. Brain Mapp.* 40, 2358–2376. doi: 10.1002/hbm.24528
- Lu, W., Dong, K., Cui, D., Jiao, Q., and Qiu, J. (2019). Quality assurance of human functional magnetic resonance imaging: a literature review. *Quant. Imaging Med. Surg.* 9, 1147162–1141162. doi: 10.21037/qims.2019.04.18
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *eLife* 10, e71774. doi: 10.7554/eLife.71774.sa2
- Mueller, K., Lepsien, J., Möller, H. E., and Lohmann, G. (2017). Commentary: cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Front. Hum. Neurosci.* 11, 345. doi: 10.3389/fnhum.2017.00345
- Muschelli, J., Nebel, M. B., Caffo, B. S., Barber, A. D., Pekar, J. J., and Mostofsky, S. H. (2014). Reduction of motion-related artifacts in resting state fMRI using aCompCor. *Neuroimage* 96, 22–35. doi: 10.1016/j.neuroimage.2014.03.028
- Nebel, M. B., Lidstone, D. E., Wang, L., Benkeser, D., Mostofsky, S. H., and Risk, B. B. (2022). Accounting for motion in resting-state fMRI: what part of the spectrum are we characterizing in autism spectrum disorder? *Neuroimage* 257, 119296. doi: 10.1016/j.neuroimage.2022.119296
- Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., et al. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U. S. A.* 89, 5951–5955. doi: 10.1073/pnas.89.13.5951
- Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., et al. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* 18, 115–126. doi: 10.1038/nrn.2016.167
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* 84, 320–341. doi: 10.1016/j.neuroimage.2013.08.048
- Power, J. D., Silver, B. M., Silverman, M. R., Ajodan, E. L., Bos, D. J., and Jones, R. M. (2019). Customized head molds reduce motion during resting state fMRI scans. *Neuroimage* 189, 141–149. doi: 10.1016/j.neuroimage.2019.01.016
- Van Dijk, K. R. A., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* 59, 431–438. doi: 10.1016/j.neuroimage.2011.07.044
- Yan, C. G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., et al. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *Neuroimage* 76, 183–201. doi: 10.1016/j.neuroimage.2013.03.004
- Yang, H., Long, X.-Y., Yang, Y., Yan, H., Zhu, C.-Z., Zhou, X.-P., et al. (2007). Amplitude of low frequency fluctuation within visual areas revealed by resting-state functional MRI. *Neuroimage* 36, 144–152. doi: 10.1016/j.neuroimage.2007.01.054
- Yeung, A. W. K., Wong, N. S. M., and Eickhoff, S. B. (2020). Empirical assessment of changing sample-characteristics in task-fMRI over two decades: an example from gustatory and food studies. *Hum. Brain Mapp.* 41, 2460–2473. doi: 10.1002/hbm.24957
- Zang, Y., Jiang, T., Lu, Y., He, Y., and Tian, L. (2004). Regional homogeneity approach to fMRI data analysis. *Neuroimage* 22, 394–400. doi: 10.1016/j.neuroimage.2003.12.030



OPEN ACCESS

EDITED BY

Richard Craig Reynolds,
Clinical Center (NIH), United States

REVIEWED BY

Can Ceritoglu,
Johns Hopkins University,
United States
Annika Carola Linke,
Western University, Canada

*CORRESPONDENCE

Céline Provins
✉ celine.provins@unil.ch
Oscar Esteban
✉ phd@oscaresteban.es

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroimaging

RECEIVED 18 October 2022

ACCEPTED 19 December 2022

PUBLISHED 12 January 2023

CITATION

Provins C, MacNicol E, Seeley SH,
Hagmann P and Esteban O (2023)
Quality control in functional MRI
studies with MRIQC and fMRIPrep.
Front. Neuroimaging 1:1073734.
doi: 10.3389/fnimg.2022.1073734

COPYRIGHT

© 2023 Provins, MacNicol, Seeley,
Hagmann and Esteban. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Quality control in functional MRI studies with MRIQC and fMRIPrep

Céline Provins^{1*}, Eilidh MacNicol², Saren H. Seeley³,
Patric Hagmann¹ and Oscar Esteban^{1*}

¹Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland, ²Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, ³Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, United States

The implementation of adequate quality assessment (QA) and quality control (QC) protocols within the magnetic resonance imaging (MRI) research workflow is resource- and time-consuming and even more so is their execution. As a result, QA/QC practices highly vary across laboratories and “MRI schools”, ranging from highly specialized knowledge spots to environments where QA/QC is considered overly onerous and costly despite evidence showing that below-standard data increase the false positive and false negative rates of the final results. Here, we demonstrate a protocol based on the visual assessment of images one-by-one with reports generated by MRIQC and fMRIPrep, for the QC of data in functional (blood-oxygen dependent-level; BOLD) MRI analyses. We particularize the proposed, open-ended scope of application to whole-brain voxel-wise analyses of BOLD to correspondingly enumerate and define the exclusion criteria applied at the QC checkpoints. We apply our protocol on a composite dataset ($n = 181$ subjects) drawn from open fMRI studies, resulting in the exclusion of 97% of the data (176 subjects). This high exclusion rate was expected because subjects were selected to showcase artifacts. We describe the artifacts and defects more commonly found in the dataset that justified exclusion. We moreover release all the materials we generated in this assessment and document all the QC decisions with the expectation of contributing to the standardization of these procedures and engaging in the discussion of QA/QC by the community.

KEYWORDS

quality control, quality assessment, fMRI, MRIQC, fMRIPrep, exclusion criteria, neuroimaging

1. Introduction

Quality assessment (QA) and quality control (QC) of magnetic resonance imaging (MRI), implemented at several stages of the processing and analysis workflow, are critical for the reliability of the results. QA focuses on ensuring the research workflow produces data of “sufficient quality” (e.g., identifying a structured artifact caused by an environmental condition that can be actioned upon so that it doesn’t replicate prospectively in future acquisitions). On the other hand, QC excludes poor-quality data from a dataset so that they do not continue through the research workflow and potentially bias results. Indeed, below-standard MRI data increase the false positive

and false negative rates in the final analyses (Power et al., 2012; Alexander-Bloch et al., 2016; Ducharme et al., 2016; Zalesky et al., 2016). For example, Power et al. (2012) showed that unaccounted-for head motion in functional MRI (fMRI) data introduces systematic but spurious spatial correlations that are wrongly interpreted as functional brain connectivity.

Despite efforts toward automation, the implementation of QA/QC checkpoints remains unstandardized and typically involves the screening of the images one by one. Therefore, QA/QC is time-consuming and frequently seen as overly onerous to the development of projects. In the absence of a consensus on systematic approaches to QA/QC and corresponding data curation protocols, laboratories currently rely on their internal know-how. Such knowledge is generally acquired through individual researchers (here, referred to as “raters”) repeatedly screening data. Thus, the knowledge is usually contingent on the context of the studies for which they are acquired and local practices rather than some principled definition of quality criteria that generalize across applications. This leads to a wide variety of QA/QC procedures and protocols across institutions, which add to the inherently large intra- and inter-rater variabilities given a specific QA/QC approach. Therefore, appropriate protocols and tools are required to make QA/QC more consistent across institutions and improve intra- and inter-rater reliability. Substantial work has been proposed to provide efficient interfaces such as MRIQC (Esteban et al., 2017), MindControl (Keshavan et al., 2018) or Swipes4Science (Keshavan et al., 2019). Large consortia have also made remarkable investments in this important task and have developed 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 for the QC of the MRI data of the UK Biobank (Alfaro-Almagro et al., 2018), where quality was defined in a more utilitarian manner as the success of downstream processing. With the rise of large-scale datasets such as the UK Biobank, manually checking the data becomes infeasible. Alfaro-Almagro et al. (2018) described an automated QC approach wherein raw data were screened for having the wrong dimensions, corrupted, missing, or otherwise unusable, and excluded from further preprocessing (first checkpoint). The second checkpoint was applying a supervised learning classifier to the T₁-weighted (T1w) images. Although image exclusions often occurred in response to qualitative issues on images (e.g., visual identification of artifacts), some images were discarded without straightforward mapping to quality issues, and the classifier was only trained to identify problems in T1w images, so it could not be applied to BOLD data or other modalities. Many researchers have similarly attempted automation, either by relying on no-reference (as no ground truth is available) image quality metrics (IQMs) to train a machine learning model (Mortamet et al., 2009; Shehzad et al., 2015; Esteban et al., 2017) or by training deep models on 3D images directly (Garcia et al., 2022). However, predicting the quality of images acquired at a

new site yet unseen by the model remains a challenging problem (Esteban et al., 2017, 2018). Another challenge to developing deep models is the need for large datasets with usable and reliable QA/QC annotations for training. Moreover, the QA/QC annotations must be acquired across sites and rated by many individuals to ensure generalizability (Keshavan et al., 2019).

Here, we demonstrate a protocol for the QC of task-based and resting-state fMRI studies. This contribution is part of the research topic “Demonstrating Quality Control (QC) Procedures in fMRI.” The participants of the research topic were given a composite dataset with anatomical and functional data selected from published studies to demonstrate QC protocols in practice. We describe how the overall application scope (that is, the intended use of the data) determines how QC is carried out and define the exclusion criteria for anatomical (T₁-weighted; T1w) and functional (blood-oxygen dependent-level; BOLD) images at two QC checkpoints accordingly. We first performed QC of the unprocessed data using the MRIQC visual reports (Esteban et al., 2017). Second, for the data that surpassed this first checkpoint, we assessed the results of minimal preprocessing using the fMRIPrep visual reports (Esteban et al., 2019). Thus, reaching a consensus on the definition of QA/QC evaluation criteria and establishing standard protocols to ascertain such criteria are the keystone toward more objective QA/QC in fMRI research.

2. Methods

2.1. Data

We used the data collection preselected by the research topic organizers to showcase examples of each exclusion criterion. The dataset gathers resting-state and task fMRI data from several open, public repositories (Biswal et al., 2010; Di Martino et al., 2014; Markiewicz et al., 2021). Therefore, the dataset is eminently multi-site and highly diverse in acquisition devices, parameters, and relevant settings. The selection criteria of datasets and subjects were not disclosed to the research topic participants. The dataset is split into two cohorts: subjects with resting-state scans and subjects with task scans. Every subject has one T1w image and one or two BOLD fMRI scans. Data were released following the Brain Imaging Data Structure (BIDS; Gorgolewski et al., 2016).

2.2. Scope of application

Considering the dataset’s characteristics, we narrowed the planned analysis’s scope to “whole-brain, voxel-wise analyses of spatially standardized task and resting-state BOLD fMRI.” Note that by “whole-brain”, we mean cortex and subcortical structures but not cerebellum because we expected those regions to fall outside of the field of view in a number of the BOLD datasets. For the implementation of such an

application, we propose our fMRI protocol (Esteban et al., 2020), which uses fMRIPrep to prepare the data for analysis. fMRIPrep was executed with default settings (for the exact description of the preprocessing see [Supplementary material](#), section 4). Therefore, data are spatially standardized into the MNI152NLin2009cAsym space (Fonov et al., 2009) accessed with TemplateFlow (Circ et al., 2022). The protocol involves an initial QC checkpoint implemented with MRIQC and a second QC checkpoint on the outputs of fMRIPrep.

2.3. QC protocol

2.3.1. Standard operating procedures (SOPs)

To formalize the scope and the QA/QC criteria and protocols, we proposed our MRIQC-SOPs template (<https://github.com/nipreps/mriqc-sops>) as a scaffold to create custom standard operating procedures (SOPs) documents tailored to the specific project and maintained under version control. We demonstrated MRIQC-SOPs to create the corresponding documentation of this study. These SOPs contain the lists presented in [Tables 1–3](#) and the QC criteria details laid out in Section 2.4 in a format adapted to the SOPs. The SOPs documents can be visualized at <http://www.axonlab.org/frontiers-qc-sops/> and can be accessed as stated in the Data and Software availability statement.

2.3.2. Image processing

Image processing was carried out according to our protocol (Esteban et al., 2020). First, we ran MRIQC with a Docker container of its latest version 22.0.1 ([Listing 1](#) shows an example script). This version performs head motion estimation with AFNI (version 22.0.17; Cox, 1996), followed by brain extraction with SynthStrip (Hoopes et al., 2022) and several image registration tasks with ANTs (version 2.3.3.dev168-g29bdf; Avants et al., 2008). Since data were already BIDS compliant, no formatting or adaptation actions were required before running MRIQC. MRIQC generated one visual report per T1w image and BOLD scan, which author CP evaluated as part of the QC protocol described below. The panels presented in the visual report are specific to the modality, meaning that different visualizations are presented for an anatomical scan compared to a functional scan. Once all the visual reports had been evaluated as indicated below (Assessment of the unprocessed data), we executed fMRIPrep only on those subjects for which the T1w and at least one BOLD scan had passed the initial QC checkpoint. As for MRIQC, fMRIPrep could be directly run on the BIDS inputs using the corresponding Docker container at version 22.0.0 (see [Supplementary material](#), section 4). As a result, fMRIPrep yielded preprocessed data and one individual QA/QC report per subject. Based on these individual reports, we established our second QC checkpoint, which was executed by author CP. The scripts we ran to execute MRIQC on the task

fMRI data and fMRIPrep on the preprocessed data can be found in the [Supplementary material](#), section 3.

2.3.3. Assessment of the unprocessed data

Visualization of reports was performed on a 27" monitor. The reports corresponding to each BOLD scan were assessed first, following the reports' ordering of visualizations. Once the full report had been visualized, CP would return to specific sections of the report when a second assessment was necessary. Finally, author CP reported her QC assessment on a spreadsheet table (included in the [Supplementary material](#)), indicating which criteria led to exclusion. The exclusion criteria are described in detail in Section 2.4. A similar protocol was then applied for screening all reports corresponding to T1w images.

2.3.4. Assessment of the minimally preprocessed data

Visualization of reports was performed on a 27" monitor. The reports corresponding to subjects that passed the previous checkpoint were screened one by one by CP. Author CP manually noted down the corresponding assessments on a spreadsheet table (included in the [Supplementary material](#)).

2.4. Assessment of quality aspects and exclusion criteria

Our exclusion criteria are all based on the visual inspection of the individual MRIQC and fMRIPrep reports, so they are all qualitative. Exclusion criteria are defined in reference to specific artifacts and qualitative aspects of BOLD and T1w images. Furthermore, we did not differentiate criteria for task and resting-state scans because our defined scope was not specific enough (e.g., lacking in objectives to determine whether some regions are of particular interest), except for the hyperintensity of single slices criterion. Each criterion is labeled for further reference in the document, the rater's notes, and the SOPs documents. [Table 1](#) exhaustively lists the exclusion criteria based on the MRIQC visual report of BOLD data, [Table 2](#) lists the criteria used to flag T1w data based on the MRIQC visual report, and [Table 3](#) lists the exclusion criteria based on fMRIPrep visual reports. These tables are also cross-referenced with each criterion's label.

2.4.1. Exclusion criteria for unprocessed BOLD data assessed with MRIQC visual reports

2.4.1.1. Artifactual structures in the background (Criterion A)

Because no BOLD signal originates from the air surrounding the head, the background should not contain visible structures. However, signals sourcing from the object of interest can spill


```

sub_nbr=$(seq 101 1 120; seq 201 1 220; seq 301 1 316; seq 401 1 420; seq 501 1 520;
seq 601 1 620; seq 701 1 720)) #subject numbers

bs=1 #batch size
for ((i=0; i<=${#sub_nbr[@]}; i+=bs)); do
    #launch mriqc on batches of subjects
    batch=${sub_nbr[@]:$i:1}
    echo ${batch[@]}
    docker run -u $( id -u ) -it --memory="8g" --rm -v
/data/datasets/QCResearchTopic/fmri-open-qc-rest/:/data:ro -v
/data/derivatives/mriqc/v22.0.1/QCResearchTopic/fmri-open-qc-rest/:/out -v
$HOME/tmp/mriqc/v22.0.1/QCResearchTopic/fmri-open-qc-rest/:/work nipreps/mriqc:22.0.1
/data /out --ica --verbose-report participant --participant-label ${batch[@]} -w /work
-vv
done

```

Listing 1

Execution of MRIQC with a Docker container. MRIQC follows the standards laid out by BIDS-Apps (Gorgolewski et al., 2017). As such, the command line using containers is composed of a preamble configuring Docker, the name of the specific Docker image (nipreps/mriqc:22.0.1), and finally, MRIQC's arguments. Because SynthStrip is a deep-learning-based approach, the brain masking step requires at least 8GB of memory (specified by the `--memory` flag).

Table 1 Resting-state and task fMRI exclusion criteria based on the MRIQC visual report.

QC of unprocessed fMRI data based on MRIQC visual report	A) Artfactual structures in the background B) Susceptibility distortion artifacts BA) Signal drop-out BB) Brain distortions C) Aliasing ghosts D) Wrap-around that overlaps with the brain E) Structured crown region in the carpet plot EA) due to motion peaks EB) due to periodic motion EC) due to coil failure ED) drift of unknown source F) Artifacts detected with independent components analysis G) Hyperintensity of single slices H) Vertical strikes in the sagittal plane of the standard deviation map I) Data formatting issues
--	---

The order of the criteria is arbitrary.

into the background through several imaging processes, e.g., aliasing ghosts, spillover originating from moving and blinking eyes, or bulkhead motion. Structures in the background are most clearly noticeable in MRIQC's "background noise panel" view, but they are frequently detectable in the standard deviation map

Table 2 T1w flagging criteria based on the MRIQC visual report.

QC of unprocessed T1w data based on the MRIQC visual report	J) Artfactual structures in the background K) Susceptibility distortion artifacts KA) Signal drop-out KB) Brain distortions L) Aliasing ghost M) Wrap-around that overlaps with the brain N) Data formatting issues O) Motion-related and Gibbs ringing P) Extreme intensity non-uniformity Q) Eye spillover
---	---

The order of the criteria is arbitrary.

view. Structure in the background is not a problem in itself as it is situated outside of the brain; the issue is that the latter artifact is likely overflowing on the brain, thus compromising brain signal. The aliasing ghost is a particular case of spurious structures in the background, discussed in further detail in criterion C below. We classified under exclusion criteria A all other structures that did not correspond to an aliasing artifact. **Supplementary Figure 1** shows several illustrative examples.

2.4.1.2. Susceptibility distortion artifacts (B)

Susceptibility distortions are caused by B_0 field non-uniformity (Hutton et al., 2002). Indeed, inserting an object in the scanner bore perturbs the nominal B_0 field, which should be constant all across the FoV. Specifically, tissue boundaries

Table 3 Resting-state and task exclusion criteria based on the fMRIPrep visual report.

QC of preprocessed data based on fMRIPrep visual report	R) Failure in normalization to MNI space S) Inaccurate brain mask T) Residual susceptibility distortion U) Error in brain tissue segmentation of T1w images V) Surface reconstruction problem W) Co-registration problem X) Regions identified for the extraction of nuisance regressors potentially cover neural signal sources
---	--

The order of the criteria is arbitrary.

produce steps of deviation from the nominal B_0 field, which are larger where the air is close to tissues. Because of these deviations, the signal is recorded at locations slightly displaced from the sampling grid along the phase encoding axis leading to susceptibility distortions (Esteban et al., 2021). Susceptibility distortions manifest in two different ways on the BOLD average panel of the MRIQC visual report (Supplementary Figure 2): as signal drop-out, that is, a region where the signal vanishes (criterion BA), or as brain distortions (criterion BB). Signal drop-outs often appear close to brain-air interfaces, as explained below; these include ventromedial prefrontal cortex, the anterior part of the prefrontal cortex, and the region next to the ear cavities. Susceptibility distortion artifacts can be corrected by the susceptibility distortion correction implemented in fMRIPrep, provided that a field map associated with the BOLD image has been acquired and is correctly referenced in the dataset. This means that the presence of susceptibility distortions does not necessarily constitute an exclusion criterion. However, given the application scope of this paper, since no field maps were shared with the dataset and because we did not identify regions of little interest where these artifacts may be less detrimental, any signal drop-out observed resulted in the exclusion of the scan. In practice, legacy datasets without field maps can still be usable if researchers take adequate mitigation approaches (which also require rigorous QA/QC).

2.4.1.3. Aliasing ghosts (C)

A ghost is a type of structured noise that appears as shifted and faintly repeated versions of the main object, usually in the phase encoding direction. They occur for several reasons, such as signal instability between pulse cycle repetitions or the particular strategy of echo-planar imaging to record the k-space during acquisition. Ghosts are often exacerbated by within-volume head motion. Sometimes they can be spotted in the BOLD average view of the MRIQC visual report, but they are more apparent in the background noise visualization. We excluded the scans for which ghosts were approximately the same intensity as the brain's interior in the background noise visualization. Supplementary Figure 3 compares an aliasing artifact that led to exclusion and one that did not.

2.4.1.4. Wrap-around (D)

Wrap-around occurs whenever the object's dimensions exceed the defined field-of-view (FOV). It is visible as a piece of the head (most often the skull, in this dataset) being folded over on the opposite extreme of the image. We excluded subjects based on the observation of a wrap-around only if the folded region contained or overlapped the cortex. In the MRIQC visual report, the wrap-around can be spotted on the BOLD average, standard deviation map, and the background noise visualization. However, we found that the background noise visualization is the clearest to assess whether the folded region overlaps the brain (Supplementary Figure 4). Note that increasing the screen's brightness helps when looking for both aliasing ghosts and wrap-around overlapping the brain, as low brightness makes the artifacts harder to see.

2.4.1.5. Assessment of time series with the carpet plot (E)

The carpet plot is a tool to visualize changes in voxel intensity throughout an fMRI scan. It works by plotting voxel time series in close spatial proximity so that the eye notes temporal coincidence (Power, 2017). Both MRIQC and fMRIPrep generate carpet plots segmented in relevant regions. One particular innovation of these carpet plots is that they contain a "crown" area corresponding to voxels located on a closed band around the brain's outer edge. As those voxels are outside the brain, we do not expect any signal there, meaning that if some signal is observed, we can interpret it as artifactual. Therefore, a strongly structured crown region in the carpet plot is a sign that artifacts are compromising the fMRI scan (Provins et al., 2022a). For example, motion peaks are generally paired with prolonged dark deflections derived from spin-history effects (criterion EA). Periodic modulations on the carpet plot indicate regular, slow motion, e.g., caused by respiration, which may also compromise the signal of interest (criterion EB). Furthermore, coil failures may be identifiable as a sudden change in overall signal intensity on the carpet plot and generally sustained through the end of the scan (criterion EC). In addition, sorting the rows (i.e., the time series) of each segment of the carpet plot such that voxels with similar BOLD dynamics appear close to one another reveals non-global structure in the signal, which is obscured when voxels are ordered randomly (Aquino et al., 2020). Thus, strongly polarized structures in the carpet plot suggest artifact influence (criterion ED). Supplementary Figure 5 illustrates the four types of carpet plot patterns. Finding temporal patterns similar in gray matter areas and simultaneously in regions of no interest (for instance, cerebrospinal fluid or the crown) indicates the presence of artifacts, typically derived from head motion. If the planned analysis specifies noise regression techniques based on information from these regions of no interest [which is standard and recommended (Ciric et al., 2017)], the risk of removing signals with neural origins is high, and affected scans should be excluded.

2.4.1.6. Artifacts detected with independent components analysis (F)

MRIQC was run with the `--ica` argument, which generates an independent component decomposition using FSL MELODIC (version 5.0.11; Beckmann and Smith, 2004). Such techniques have been thoroughly described elsewhere (Griffanti et al., 2017). Components are easily screened with the specific visualization “ICA components” in the corresponding BOLD report, and each component is mapped on a glass brain with an indication of their frequency spectrum and their corresponding weight over time. One recurring artifactual family of components emerges when motion interacts with interleaved acquisition giving rise to the so-called spin-history effects. The spin-history effects appear as parallel stripes covering the whole brain in one direction (see [Supplementary Figure 6](#)). They are a consequence of the repetition time not being much larger than the T1 relaxation time in typical fMRI designs. This implies that the spins will not completely relax when the next acquisition starts.¹ In addition, specific movements (e.g., rotation around one imaging axis, such as nodding) will exacerbate spin-history effects as slices will cut through the brain at different locations between consecutive BOLD time points. These two considerations combined mean that motion will produce spins with different excitation histories, and thus, the signal intensity will differ. Components showcasing parallel stripes concurring with slices in extreme poles of the brain or even across the whole brain are likely to capture these effects.

2.4.1.7. Hyperintensity of single slices (G)

Above the carpet plot, MRIQC and fMRIPrep represent several time series to support the interpretation of the carpet. In particular, the slice-wise z-standardized signal average is useful for detecting sudden “spikes” in the average intensity of single slices of BOLD scans. When paired with the motion traces, it is possible to determine whether these spikes are caused by motion or by possible problems with the scanner (e.g., white-pixel noise). Spikes caused by motion typically affect several or all slices, while spikes caused by white-pixel noise affect only one slice and are generally more acute (see [Supplementary Figure 7](#)). White-pixel noise is generally caused by some small pieces of metal in the scan room or a loose screw on the scanner that accumulates energy and then discharges randomly. This creates broad-band RF noise at some point during the signal read-out, leading to one spot in the k-space with abnormally high intensity. In the image domain, it manifests as an abrupt signal intensity change in one slice at one time point. The problem is particularly acute for EPI scans because of all the gradient blipping during the read-out. For resting-state data, we discarded BOLD scans containing these spikes regardless of their physical origin (motion vs. white-pixel noise) because correlation analyses are likely biased by such peaks. Conversely,

task data analyses are typically more robust to this particular artifact. Therefore the presence of only one or more relatively small spikes led to the scan being flagged for careful inspection after the preprocessing.

2.4.1.8. Vertical strikes in the sagittal plane of the standard deviation map (H)

The sagittal view of the standard deviation map might show vertical strike patterns that extend hyperintensities through the whole sagittal plane (see [Supplementary Figure 8](#)). We excluded all images showcasing these patterns. Although we did not find an explanation of the mechanism behind this artifact, email conversations dating from 2017 seemed to point at an interaction between physiology and environmental issues in the scanning room that may affect the receiver coils.

2.4.1.9. Data formatting issues (G)

As part of the NIfTI format (Cox et al., 2004), the file header contains metadata storing several relevant parameters, of which the orientation information is critical for the interpretability of the data. The orientation parameters indicate how the data matrix is stored on disk and enable visualizing rows and slices at the correct locations (Glen et al., 2020). However, mistakes may occur while recording information at the scanner, converting DICOM to NIFTI format, or during a subsequent processing step. Such mistakes result in the brain image not being correctly visualized and preprocessed, with axes either being flipped (e.g., the anterior part of the brain is labeled as posterior) or switched (e.g., axial slices are interpreted as coronal ones). These issues may render the dataset unusable, e.g., if the orientation information describing whether the data array has been recorded from left to right or right to left is lost. Examples are shown in [Supplementary Figure 9](#).

2.4.2. Criteria for flagging unprocessed T1w data based on the MRIQC visual report

Given our planned analysis, the T1w image will be used solely to guide the spatial alignment to the standard MNI152NLin2009cAsym template. In addition, surface reconstructions from the T1w image will guide the co-registration of structural and functional (BOLD) images in fMRIPrep. Since the latter preprocessing steps are relatively robust to structural images with mild artifacts, we did not impose exclusion criteria on the unprocessed T1w images. However, we annotated subjects with visible artifacts in the T1w images to ensure rigorous scrutinizing of spatial normalization and surface reconstruction outputs from fMRIPrep (if both modalities passed the first QC checkpoint with MRIQC). The explanation and the description of the criteria J to N are the same as their counterpart in Section 2.4.1 and are illustrated in [Supplementary Figure 10](#).

¹ <https://imaging.mrc-cbu.cam.ac.uk/imaging/CommonArtefacts>

2.4.2.1. Motion-related and Gibbs ringing (O)

Large head motion during the acquisition of T1w images often expresses itself with the appearance of concentric ripples throughout the scan (see [Supplementary Figure 10E](#)). In the most subtle cases, motion-related ripples look similar to the fine lines generated by Gibbs ringing. The latter emerges as a consequence of the truncation of the Fourier series approximation and appears as multiple fine lines immediately adjacent and parallel to high-contrast interfaces. While Gibbs ringing is limited to the adjacency of sharp steps in intensity at tissue interfaces, the ripples caused by motion generally span the whole brain and are primarily visible in the sagittal view of MRIQC's mosaic views.

2.4.2.2. Intensity non-uniformity (P)

Intensity non-uniformity is characterized by a smooth variation (low spatial frequency) of intensity throughout the brain caused by the stronger signal sensed in the proximity of coils. It is noticeable on the zoomed-in view on the T1w image (see [Supplementary Figure 10F](#)). Furthermore, intensity non-uniformity can be a problem for automated processing methods that assume a type of tissue [e.g., white matter (WM)] is represented by voxels of similar intensities across the whole brain. An extreme intensity non-uniformity can also be a sign of coil failure.

2.4.2.3. Eye spillover (Q)

Eye movements may trigger the signal leakage from the eyes through the imaging axis with the lowest bandwidth (i.e., acquired faster), potentially overlapping signal from brain tissue. On data preserving facial features, the streak of noise is visible in the background at the levels of the eyes. However, because all the data in this study are openly shared after defacing (for privacy protection reasons), the signal around the face has been zeroed, leading to this leakage not being visible ([Provins et al., 2022b](#)). A strong signal leakage can, however, be noticeable on the zoomed-in view of the T1w image (see [Supplementary Figure 10G](#) for an example of the latter case).

2.4.3. Exclusion criteria of pre-processed data based on fMRIPrep visual report

2.4.3.1. Failure in normalization to MNI space (R)

Because the conclusions of the hypothetical analysis are based on data normalized to a standard template, the normalization must be successful. The fMRIPrep report contains a widget to assess the quality of the normalization to MNI space. The widget flickers between the MNI template and the individual's T1w image normalized to that template. To verify successful normalization, we assessed the correct alignment of the following structures (in order of importance): (1) ventricles, (2) subcortical regions, (3) corpus callosum, (4) cerebellum, and (5) cortical gray matter (GM). A misalignment of the

ventricles, the subcortical regions, or the corpus callosum led to immediate exclusion. However, we were more lenient with the misalignment of cortical GM because volumetric (image) registration may not resolve substantial inter-individual differences (e.g., a sulcus missing in an individual's brain but typically present in the population of the template). Any extreme stretching or distortion of the T1w image also indicates a failed normalization.

2.4.3.2. Inaccurate brain mask (S)

The brain mask computed from the T1w image is shown in the "brain mask and brain tissue segmentation of the T1w" panel under the anatomical section of the fMRIPrep visual report. The latter should closely follow the contour of the brain. An inaccurate brain mask presents "bumps" surrounding high-intensity areas of signal outside of the cortex (e.g., a mask including patches of the skull) and/or holes surrounding signal drop-out regions. Having an accurate brain mask makes the downstream preprocessing of an fMRI scan faster (excluding voxels of non-interest) and more accurate (less bias from voxels of non-interest). Consequently, it is important to discard subjects for which the brain mask is not well defined. Note that the brain mask plotted in the "brain mask and (anatomical/temporal) CompCor ROIs" panel under the functional section is not identical to the brain mask mentioned in this paragraph, as it is computed from the BOLD image. This mask must not leave out any brain area. Therefore, an exclusion criterion can be established when the mask intersects brain-originating signals.

2.4.3.3. Residual susceptibility distortion (T)

For cases that were not excluded following criterion B, susceptibility distortions were evaluated with the fMRIPrep report after preprocessing. Any observation of susceptibility distortion artifacts led to the exclusion of the scan (see [Supplementary Figure 11](#)).

2.4.3.4. Error in brain tissue segmentation of T1w images (U)

The panel "brain mask and brain tissue segmentation of the T1w" under the anatomical section of the fMRIPrep report shows contours delineating brain tissue segmentations overlaid on the T1w image. To confirm the good quality of the segmentation, we first verified that the pink contour accurately outlined the ventricles, whereas the blue contour followed the boundary between GM and WM. The first exclusion criterion was thus the inclusion of tissues other than the tissue of interest in the contour delineations. T1w scans showcasing a low signal-to-noise ratio because of thermal noise will present scattered misclassified voxels within piecewise-smooth regions (generally more identifiable in the WM and inside the ventricles). These scans were excluded except for images where these voxels are only present at subcortical structures (e.g., thalamus) or nearby tissue boundaries. In the

latter case, the misclassification results from partial volume effects (i.e., indeed, such voxels contain varying fractions of two or more tissues). [Supplementary Figure 12](#) illustrates the difference between individual dots caused by noise vs. partial volume effects.

2.4.3.5. Surface reconstruction problem (V)

The WM surface (blue outline) and the pial surface (red outline) reconstructed with FreeSurfer [version 7.0.1, [Fischl, 2012](#)] are overlaid on the participant's T1w image, in the panel dedicated to surface reconstruction visualization under the anatomical section of the fMRIPrep report. Since the cerebellum and the brainstem are excluded from the surface reconstruction, the outlines will not include these areas. QC assessment of FreeSurfer outcomes is comprehensively covered elsewhere (e.g., [White et al., 2018](#); [Klapwijk et al., 2019](#)), and fMRI studies using vertex-wise (surface) analyses should rigorously assess these surfaces. In this protocol, we only excluded data when the reconstructed surfaces were extremely inaccurate, which typically only happens in the presence of artifacts easily captured previously by MRIQC (Section 2.4.2).

2.4.3.6. Co-registration problem (W)

The fMRIPrep report contains a widget to assess the accuracy of the alignment of BOLD runs into the individual's anatomical reference (co-registration). The widget flickers between the reference T1w image and the BOLD average co-registered onto it. Extracted brain surfaces' contours are represented as further anatomical cues. Here, we checked the alignment of image intensity edges and the anatomical landmarks (e.g., the ventricles and the corpus callosum) between the BOLD and the T1w images.

2.4.3.7. Regions identified for the extraction of nuisance regressors potentially cover neural signal sources (X)

fMRIPrep calculates CompCor ([Behzadi et al., 2007](#)) nuisance regression time series to remove physiological and head motion artifacts from BOLD scans. Two families of CompCor methodologies are provided within the outputs: temporal CompCor (tCompCor) uses voxels presenting the highest temporal variability, and anatomical CompCor (aCompCor) extracts signal from regions of no interest (e.g., a conservative mask including core areas of the ventricles and the WM). fMRIPrep provides a panel to assess the adequacy of these regions from which CompCor will extract regressors ("brain mask and anatomical/temporal CompCor ROIs"). In addition to the masks corresponding to CompCor, the "crown" mask can also be assessed in this visualization. If the study plan prescribes using CompCor or brain-edge regressors, it is critical to exclude BOLD runs where any of these masks substantially overlap regions of interest.

3. Results

Following our predefined exclusion criteria, we excluded all the BOLD scans at the first QC checkpoint, except 4/151 for the resting-state subset and 1/30 for the task subset (97% of the subjects were excluded). The high exclusion rate was expected as this dataset had been preselected to contain data expressing a wide range of artifacts. In a standard dataset, the exclusion rate usually lays between 10 and 25% ([Esteban et al., 2017](#)). By far, the most common reason for exclusion was the presence of susceptibility distortion (exclusion criterion B). Other commonly found artifacts that met the exclusion criteria included aliasing ghost (C), problematic wrap-around (D), and structured carpet plots (E). The number of times each criterion has been cited as a reason for exclusion is reported in [Supplementary Table 1](#). Moreover, 58/181 T1w images were flagged for thorough scrutinization of the normalization and the surface reconstruction outputs of fMRIPrep. One T1w image was exceptionally excluded based on the MRIQC visual report because of extreme motion-related ringing. An overview of how often a scan was flagged based on which criterion can be found in [Supplementary Table 2](#). Out of the five subjects that passed the first QC checkpoint, two were excluded based on the inspection of the fMRIPrep visual reports for the presence of previously undetected signal drop-out. Some of our criteria did not result in the exclusion of data in this dataset: spin-history effects, failed normalization, problematic brain masks of either T1w or BOLD images, surface reconstruction problem, and failed co-registration.

3.1. QC of MRI data substantially relies on the background

The visual assessment of the "background noise" section of MRIQC reports helps unveil several artifactual structures suggesting further issues within the regions of interest (see [Figure 1A](#)). Aliasing ghosts that manifest as faint and shifted copies of the brain visible in the background are a particular type of structure in the background (see [Figure 1B](#)). Secondly, the background enclosed by the crown region plays an important role in detecting structure in the carpet plot. The influence of motion outbursts can be seen as prolonged dark deflections (see [Figure 1C](#)). Conversely, the presence of periodic modulation of the intensity can be attributed to periodic motion related to respiration (see [Figure 1D](#)). Thirdly, following the assumption that the slice-wise noise average on the background should be smooth, peaks in the single slices indicate some issue at the acquisition (i.e., white-pixel noise illustrated in [Figure 1E](#)). Overall, in adult MRI, no BOLD signal originates from the background, meaning that structures visible in the background

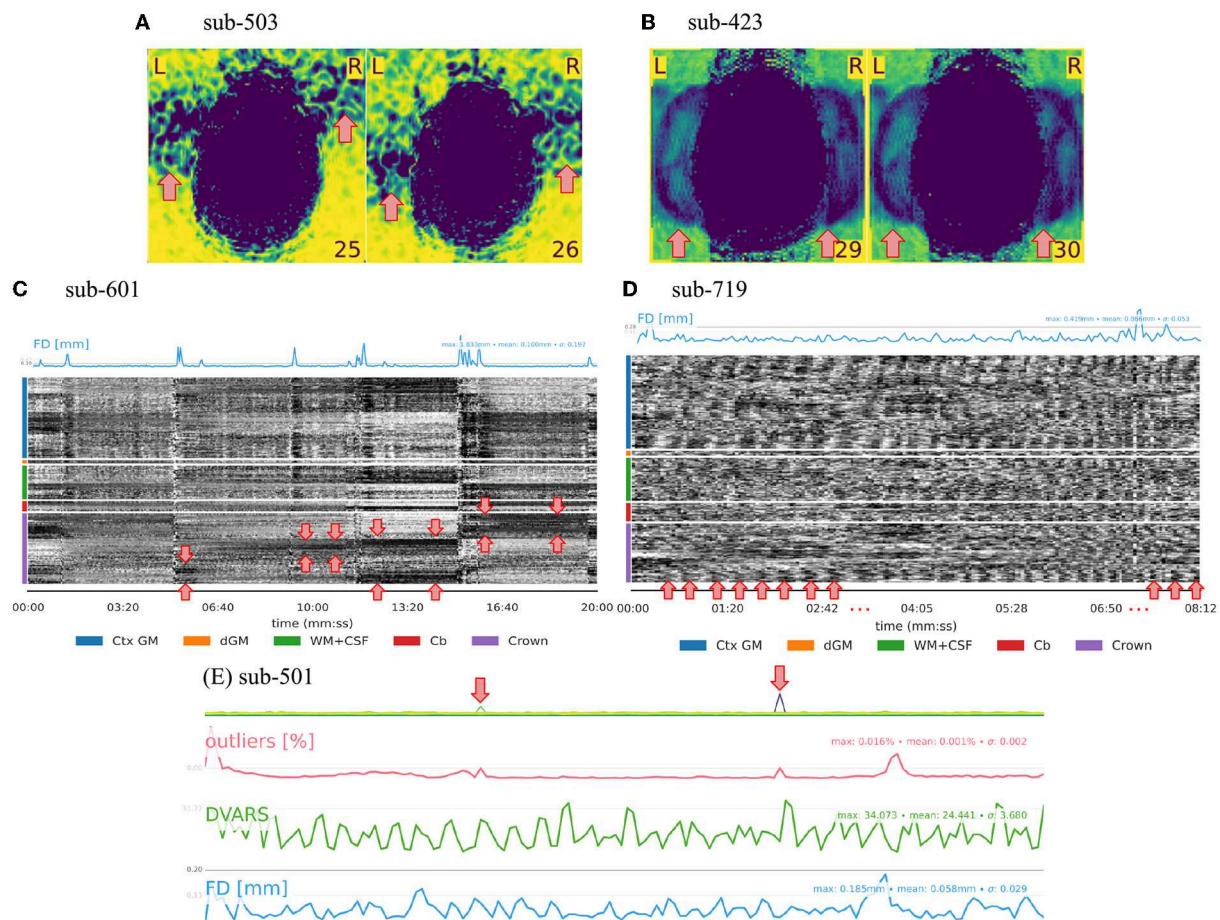


FIGURE 1

QA/QC of MRI data relies substantially on the background. Several exclusion criteria listed in Tables 1, 2 are based on the background. (A) Heavy structure in the background constitutes an exclusion criterion as the artifact likely extends inside the brain thus compromising signals of interest. (B) Aliasing ghosts appear as a faint and shifted copy of the brain in the background. (C, D) Since the crown comprises voxels outside the brain, the structure in the crown region of the carpet plot springs from artifacts. For example, two types of motion-related patterns can be distinguished. (C) Prolonged dark deflections are often caused by motion outbursts, visible as peaks in the framewise displacement (FD) trace. (D) Periodic fluctuations of intensity throughout the carpet plot can be attributed to periodic motion due to respiration. (E) The presence of sudden intensity change in a single slice can be attributed to white-pixel noise and constitutes an exclusion criterion.

come from artifacts. This consideration renders the background a convenient resource to assess MRI scans.

3.2. Setting QC checkpoints at several steps of the preprocessing is important

In this protocol, we illustrate how we set up two QC checkpoints: one for unprocessed data using MRIQC visual report and one for minimally preprocessed data using fMRIPrep visual report. Only the data that survived the first QC checkpoint with MRIQC were run through fMRIPrep, illustrating how QC must drop data that meet exclusion criteria. The checkpoint leveraging fMRIPrep's visual report is important not only to capture problems in the processing of the data (e.g., failure in co-registration) but it also offers another opportunity to look at

the data from different perspectives. To illustrate this point, we simulated a scenario where exclusion criteria were intendedly misapplied in the QC checkpoint based on MRIQC for one subject (sub-408), and as a result the dataset was inappropriately run through fMRIPrep. Figure 2A presents the tCompCor mask obtained for this subject, which suggests the presence of an artifact by its shape and its large overlap with the region of interest. These considerations justified the exclusion of the subject. Note furthermore that we did not detect that specific artifact in the MRIQC visual report (even after specifically looking out for it), illustrating the value of looking at the data using many different visualizations. Besides, viewing many slices cutting in several planes helps to not overlook exclusion criteria as illustrated in Figures 2B, C. Indeed, a signal drop-out that appeared very clearly on a specific sagittal slice (see Figure 2B) was more subtle to detect on axial slices (see Figure 2C). This

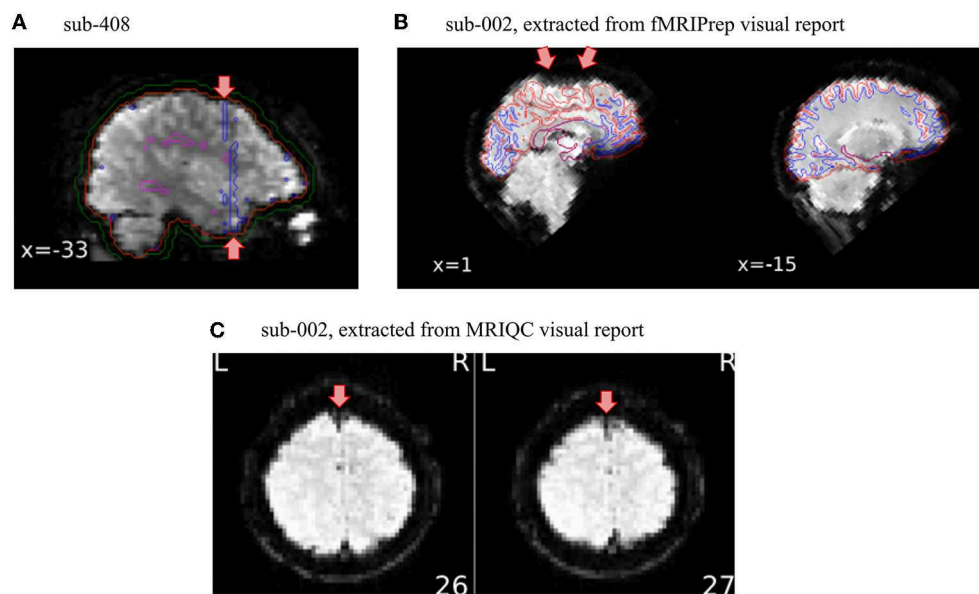


FIGURE 2

Setting QA/QC checkpoints at several steps of the preprocessing is important. Overlooking exclusion criteria while inspecting the visual reports can happen. Thus, having several QA/QC checkpoints set up along the preprocessing pipeline is valuable to catch those missed substandard scans. (A) In this particular case, the shape of the tCompCor mask looks suspiciously induced by an artifact, which led us to exclude this subject from further analysis. (B) This sagittal slice of the BOLD average presented in the fMRIPrep visual report clearly shows susceptibility distortion on the superior frontal cortex. This specific slice however did not appear in the MRIQC visual report. (C) The signal drop-out was furthermore more subtle on the axial slices, leading to an overlook of this artifact on the QA/QC checkpoint of unprocessed data.

specific sagittal BOLD average slice was displayed in the panel “Alignment of functional and anatomical MRI data (surface driven)” of the functional part of the fMRIPrep report, a panel for which the original purpose is to assess the quality of co-registration and not to visualize BOLD average. This reinforces again the importance of viewing the data from different perspectives.

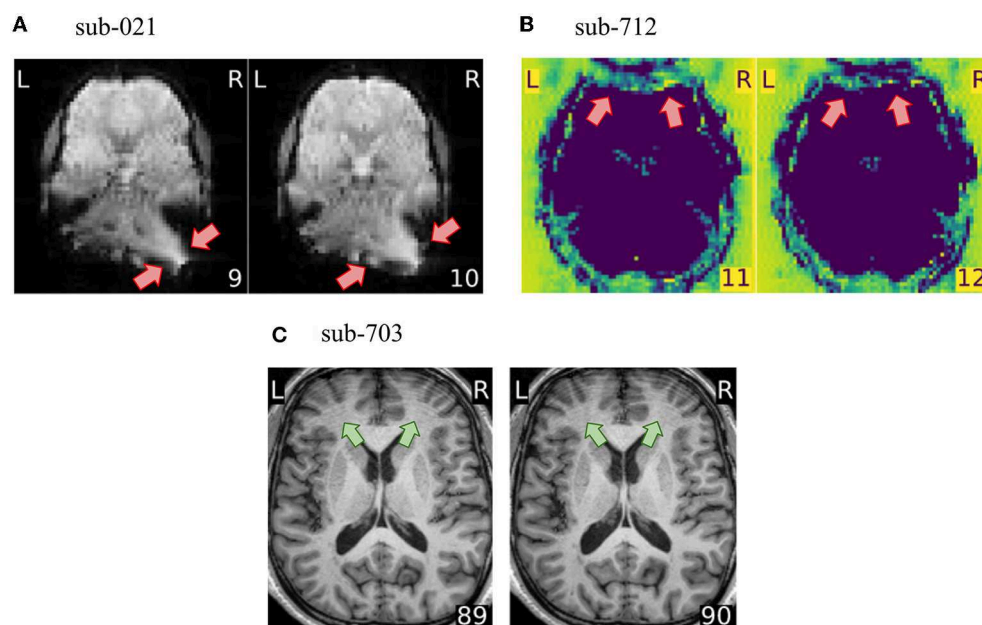
3.3. Exclusion criteria depend on the particularities of the project

How QA/QC is performed must be defined in close relation to the scope, goals, and approach of the project at hand. The first consideration is the types of data available. For example, the absence of field maps in the dataset led us to exclude a substantial portion of subjects that presented susceptibility distortion artifacts (see Figure 3A). Susceptibility distortion artifacts not only appeared as signal drop-outs or brain distortions, but they also interacted with head motion creating ripples that blurred the structure and destroyed contrast (see Supplementary Figure 2E). If field maps had been acquired and shared with the dataset, such artifacts could have been corrected by the susceptibility distortion correction run by fMRIPrep. This means that distortion present in the preprocessed data would grant exclusion at the corresponding checkpoint, but distortion

present in the unprocessed data should not be considered an exclusion criterion. This example also highlights the importance of defining the exclusion criteria according to the placement of the QA/QC checkpoint within the research workflow. A further consideration is that the research question informs the regions where quality is most important. In the hypothetical scenario that a study investigates functional activity in the motor cortex, a wrap-around that affects the prefrontal cortex (see Figure 3B) would unlikely bias analyses limited to the region of interest. As such, it would not be considered an exclusion criterion in a study about the motor cortex. On the contrary, it would be very problematic for a study focusing on, e.g., decision-making. Finally, the planned analysis also determines the implementation of QA/QC protocols. In this paper, we did not exclude T1w images presenting motion-related ringing (see Figure 3C) because the application was scoped as a functional, voxel-wise analysis. If, instead, we would have set the application’s scope as a vertex-wise (surface) analysis, then ringing on the T1w image would have granted exclusion, as the reconstructed brain surfaces from T1w images presenting the artifact would have been unreliable.

4. Discussion

We presented a QC protocol implemented on top of our previous fMRI analysis protocol (Esteban et al., 2020). We

**FIGURE 3**

The exclusion criteria depend on the particularities of the project. **(A)** fMRIPrep can correct for susceptibility distortions when field maps are available. In this project, we however consider susceptibility distortion artifacts as exclusion criteria because no field maps were shared with the dataset. **(B)** A wrap-around overlapping the prefrontal cortex would not necessarily yield scan exclusion if the research question would focus on e.g., motor cortex. Our application scope has been defined as voxel-wise whole-brain fMRI analysis, thus this wrap-around is problematic. **(C)** Motion-related ringing on the T1w image does not constitute an exclusion criterion in our protocol, because the T1w is used solely for guiding the normalization and the co-registration. However, if the application scope would use surface-based analysis, this ringing would distort surface reconstruction.

further restricted the scope of the planned analyses within standard whole-brain, voxel-wise models for both task and resting-state fMRI. Under such delineation of the application, we proposed two QC checkpoints: first, on the unprocessed data with MRIQC, and second, on the minimally preprocessed data with fMRIPrep. To fully reflect best practices, we only preprocessed the data corresponding to subjects for which the T1w image and at least one BOLD run had passed the first QC checkpoint. In this report, we also described the exclusion criteria that we believe would match the planned application and clearly remark that it is critical that researchers define these exclusion criteria in the most comprehensive way before the data are acquired (or accessed, in case of reusing existing data).

Here, we also restricted our protocol to describe QC decisions (i.e., excluding sub-standard data that risk biasing the final results). We did not describe relevant QA aspects and actions that can be triggered by QC outcomes because all data in the study were reused. Indeed, the outputs of QC should be leveraged to prevent quality issues from propagating through prospective acquisition. One example of how QA is limited in studies reusing data is the availability of field maps

to correct susceptibility-derived distortions in BOLD images. Indeed, when field maps are available, fMRIPrep will run susceptibility distortion correction by default. However, no field maps were available in the dataset. Although we could have used fMRIPrep's "field map-less" approach to address susceptibility distortions, we decided such a decision would complicate the QC protocol description with an experimental, non-standard feature of fMRIPrep. A second QA aspect derived from the example dataset is the choice of the phase encoding direction. The phase encoding direction is generally the most limited in terms of bandwidth, and as a result, most artifacts propagate along that direction. For example, in the case of eye spillover, eye movements are likely to produce artifacts, thus selecting the phase encoding to occur along the anterior-to-posterior direction over left-to-right will produce a larger overlap of artifacts with the brain. In practice, if a particular task involves eye motion (e.g., blinking, saccade), the left-right direction could be the better choice if no other consideration conflicts regarding phase encoding.

One often overseen aspect of QA/QC protocols is establishing strategies to account for raters' reliability. Indeed,

intra-rater variability and drifts are strongly driven by the protocol implementation settings (e.g., changing the size of the screen or other screen technical capabilities), training, and attrition. Raters' training is particularly relevant, and it originates systematic differences in how QA/QC criteria are applied over the time span of the project. Therefore, it is critical to use mitigation strategies like randomly selecting a few earlier reports for re-evaluation or annotating subjects one is uncertain about and returning to it later in the QC process. Similarly, the implementation of QA/QC protocols must also plan for multiple raters and anticipate a plan to counter inter-rater variabilities and drifts idiosyncratic to each of them (e.g., defining a training program with specific examples, inter-rater "calibration", etc.). Learned insights can be transferred in several ways: 1. from other subjects that expressed the artifact more clearly, 2. from examining the report of another modality of the same subject, 3. from a more senior rater, or 4. from visual inspection of other tools' output. For example, if the brain is not perfectly aligned with the imaging axes, the space between the cerebellum and the temporal lobe at the basal part of the brain appears bigger on one side of the other on axial slices. Inexperienced raters may interpret that some artifact occurred, although, in fact, the image is just visualized with some obliquity with respect to the sagittal plane. This misinterpretation would be more likely for BOLD images, as this might look like a single-sided signal drop-out.

A fundamental aspect of a robust QC protocol we have showcased is its funneling design. At every QC checkpoint, we must pre-establish some exclusion criteria that will result in dropping sub-standard data. For datasets limited in sample size, excluding data may reduce the power of the study below the planned estimation. More generally, even when the analysis plan anticipated some data replacement measures for data dropped at the earlier QC checkpoints, excluding data increases the costs of the study (in terms of scan time, subject time, etc.). In this case, real-time QA/QC (that is, during the acquisition or immediately after) is a promising strategy to minimize data exclusion and replacement costs (Heunis et al., 2020). Therefore, establishing these criteria will present the researcher with the challenge of striking an appropriate balance between being excessively stringent (and therefore, discarding too many images) and too lenient to the point that results are not reliable. For this reason, it is important to establish QC criteria from the perspective of all the QC checkpoints in the pipeline and to ensure the best trade-off. When developing this manuscript, we understood that setting the scope to "whole-brain voxel-wise" analyses would allow more flexible QC criteria for the T1w images at the MRIQC step and only mark borderline images for a more rigorous screening after the second QC checkpoint. Conversely, we also discovered some artifacts in the fMRIPrep visual report that could have been spotted in the MRIQC visual

report of the same participant. Going back to the MRIQC visual report, we could understand why this detail escaped us at the first iteration and learn from our mistake. Therefore, layering QC checkpoints is critical to ensure the robustness of the whole protocol.

4.1. Limitations and deviations from our standard QC protocols

Several limitations stem from the specifics of the dataset used in this study. First, we could not take advantage of the MRIQC group report, in which the IQMs extracted from all images in a dataset are presented in scatter plots, because this dataset was composed from multiple sources, which makes these reports hard to interpret without "harmonizing" the IQMs. On a single-site dataset, we would use the MRIQC group report to spot outliers in the IQMs distributions and double-check their corresponding visual reports for exclusion criteria. Second, we used the same exclusion criteria for the resting-state and task fMRI data, with the exception of criterion G (hyperintensity of single slices). In this particular case, we excluded resting-state runs showcasing G because this artifact will likely introduce correlations in the data that will potentially be interpreted as functional connections in such analyses. Conversely, models typically applied for analyzing task paradigms are generally more resilient to biases introduced by these hyperintensities. Third, the quality of the T1w images may have been overestimated because the data are defaced. As we explored in a recent pilot study (Provins et al., 2022b) defacing, though necessary to protect participants' privacy when sharing data publicly, likely biases manual QA/QC of anatomical images. One of our conclusions was that defaced images were perceived as having a better quality overall. Fourth, as a result of the QC data funnel mentioned above, the number of subjects for which we assessed the visual reports of fMRIPrep was severely limited to only the five out of 181 that passed the first QC checkpoint with MRIQC. The number of subjects successfully passing the first checkpoint would have been much higher if available field maps had been available within each subject's data, considering that criterion B (susceptibility distortions) was by large the top one criterion that granted exclusion of images. Lastly, the scope of the study was limited to describing the protocols and communicating our assessments. Although it would have been of interest to evaluate inter-rater and intra-rater variabilities, the settings were not adequate to address such questions. Indeed, with such a high (and expected) exclusion rate, in addition to the task of identifying as many subpar images as possible, both sources of variability in quality annotations will be certainly minimal. We explored such variabilities in Provins et al. (2022b) and we are currently extending the study with the pre-registration of a larger scale analysis (Provins et al., 2022c).

5. Conclusion

Establishing appropriate QC protocols adds to the list of practices conducive toward reliable neuroimaging workflows. Moreover, standardizing these protocols is critical to minimize intra-, and inter-rater, as well as intra- and inter-laboratories variabilities, thereby achieving consensus regarding QA/QC across researchers and opening ways to consistently train machine agents to automate the process. Therefore, the research topic in which this work is framed is a timely initiative pursuing such goals. We demonstrated the implementation of a QC protocol in a standard functional MRI analysis workflow at two checkpoints: (i) assessing the unprocessed data (with MRIQC) and (ii) assessing minimally preprocessed data (with fMRIPrep). We expect this thorough description of the QC protocol and associated data exclusion criteria built upon this research topic's initiative to promote best practices in QA/QC and help researchers implement their protocols for functional MRI more effectively.

Data availability statement

The data used in this study are publicly available at open-source repositories under permissive licenses, and a single collection with all the original data is accessible at <https://osf.io/qaesm/> under the CC-BY Attribution 4.0 International license. All MRIQC's visual reports referenced throughout the manuscript are openly available under the terms of the CC0 license at <https://mriqc.s3.amazonaws.com/index.html#FrontiersQC/>. Correspondingly, all fMRIPrep's visual reports are openly available under the CC0 license at <https://fmriprep.s3.amazonaws.com/index.html#FrontiersQC/>. The SOPs repository generated from the MRIQC-SOPs template can be found at <https://github.com/frontiers-qc-sops> and is distributed under a CC0 license. Correspondingly, a human-readable documentation website is generated with every update to the SOPs, accessible at <http://www.axonlab.org/frontiers-qc-sops/>. For archival purposes, the exact version that accompanies this manuscript is available with <https://doi.org/10.5281/zenodo.7221547>.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

Conceptualization, funding acquisition, project administration, and supervision: OE. Data curation and visualization: CP. Methodology and writing: CP, OE, and SS. Resources: PH and OE. All authors reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work has been supported by the NIMH (RF1MH121867, OE). CP and OE receive support from the Swiss National Science Foundation—SNSF—(#185872, OE). PH receives support from SNSF (#185897, PH). SS receives support from NIMH under award #T32MH122394.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnimg.2022.1073734/full#supplementary-material>

References

- Alexander-Bloch, A., Liv Clasen, M. S., Lisa Ronan, F. L., and Jay Giedd, A. R. (2016). Subtle in-scanner motion biases automated measurement of brain anatomy from *in vivo* MRI. *Hum. Brain Map.* 37, 2385–2397. doi: 10.1002/hbm.23180
- Alfaro-Almagro, F., Mark Jenkinson, N. K. B., Jesper, L. R., Andersson, L. G., and Gwenaëlle Douaud, S. N. S. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034

- Aquino, K. M., Ben, D., Fulcher, L. P., and Kristina Sabarodien, A. F. (2020). Identifying and removing widespread signal deflections from fMRI data: rethinking the global signal regression problem. *NeuroImage* 212, 116614. doi: 10.1016/j.neuroimage.2020.116614
- Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004
- Beckmann, C. F., and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imag.* 23, 137–152. doi: 10.1109/TMI.2003.822821
- Behzadi, Y., Khaled Restom, J. L., and Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* 37, 90–101. doi: 10.1016/j.neuroimage.2007.04.042
- Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Ciric, R., Thompson, W. H., Lorenz, R., Goncalves, M., MacNicol, E. E., Markiewicz, C. J., et al. (2022). TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nat. Meth.* 19, 1568–1571. doi: 10.1038/s41592-022-01681-2
- Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage* 154, 174–187. doi: 10.1016/j.neuroimage.2017.03.020
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Cox, R. W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., and Holmes, C. J. (2004). “A (Sort of) new image data format standard: NIFTI-1,” in *10th Annual Meeting of the Organization for Human Brain Mapping* (Budapest).
- Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Ducharme, S., Albaugh, M. D., Nguyen, T. V., Hudziak, J. J., Mateos-Pérez, J. M., Labbe, A., et al. (2021). Trajectories of cortical thickness maturation in normal brain development—the importance of quality control procedures. *NeuroImage* 125, 267–279. doi: 10.1016/j.neuroimage.2015.10.010
- Esteban, O., Adebimpe, A., Markiewicz, C. J., Goncalves, M., Blair, R. W., Cieslak, M., et al. (2021). “The bermuda triangle of D- and f-MRI sailors—software for susceptibility distortions,” in *27th Annual Meeting of the Organization for Human Brain Mapping (OHBM)*. p. 1653. doi: 10.31219/0sf.io/g8nt
- Esteban, O., Ciric, R., Finc, K., Blair, R. W., Markiewicz, C. J., Moodie, C. A., et al. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nat. Prot.* 15, 2186–2202. doi: 10.1038/s41596-020-0327-3
- Esteban, O., Daniel Birman, M. S., Oluwasanmi, O., Koyejo, R. A. P., and Gorgolewski, K. J. (2017). MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE* 12, e0184661–e0184661. doi: 10.1371/journal.pone.0184661
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Meth.* 16, 111–116. doi: 10.1038/s41592-018-0235-4
- Esteban, O., Poldrack, R. A., and Gorgolewski, K. J. (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. Lect. Notes Comput. Sci.* p. 190–99. doi: 10.1007/978-3-030-01364-6_21
- Fischl, B. (2012). FreeSurfer. *NeuroImage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021
- Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., and Collins, D. L. (2009). Unbiased non-linear average age-appropriate brain templates from birth to adulthood. *NeuroImage* 47, S102. doi: 10.1016/S1053-8119(09)70884-5
- Garcia, M., Dosenbach, N., and Kelly, C. (2022). BrainQCNet: a deep learning attention-based model for multi-scale detection of artifacts in brain structural MRI scans. *bioRxiv*. doi: 10.1101/2022.03.11.483983
- Glen, D. R., Taylor, P. A., Buchsbaum, B. R., Cox, R. W., and Reynolds, R. C. (2020). Beware (Surprisingly Common) left-right flips in your MRI data: an efficient and robust method to check MRI dataset consistency using AFNI. *Front. Neuroinform.* 14, 18. doi: 10.3389/fninf.2020.00018
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3, 160044. doi: 10.1038/sdata.2016.44
- Gorgolewski, K. J., Fidel Alfaro-Almagro, T. A., Pierre Bellec, M. C., and Mallar Chakravarty, M. N. W. C. (2017). BIDS apps: improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Comp. Bio.* 13, e1005209. doi: 10.1371/journal.pcbi.1005209
- Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M. F., et al. (2017). Hand classification of fMRI ICA noise components. *NeuroImage* 154, 188–205. doi: 10.1016/j.neuroimage.2016.12.036
- Heunis, S., Rolf Lamerichs, S. Z., Cesar Caballero-Gaudes, J. F. A. J., and Bert Aldenkamp, M. B. (2020). Quality and denoising in real-time functional magnetic resonance imaging neurofeedback: a methods review. *Hum. Brain Map.* 41, 3439–3467. doi: 10.1002/hbm.25010
- Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B., and Hoffman, M. (2022). SynthStrip: skull-stripping for any brain image. *NeuroImage* 260, 119474. doi: 10.1016/j.neuroimage.2022.119474
- Hutton, C., Andreas Bork, O. J., Ralf Deichmann, J. A., and Turner, R. (2002). Image distortion correction in fMRI: a quantitative evaluation. *NeuroImage* 16, 217–240. doi: 10.1006/nimg.2001.1054
- Keshavan, A., Esha Datta, I. M. M., Christopher, R., Madan, K. J., and Henry, R. G. (2018). Mindcontrol: a web application for brain segmentation quality control. *NeuroImage* 170, 365–372. doi: 10.1016/j.neuroimage.2017.03.055
- Keshavan, A., Jason, D., and Yeatman, A. R. (2019). Combining citizen science and deep learning to amplify expertise in neuroimaging. *Front. Neuroinform.* 13, 29. doi: 10.3389/fninf.2019.00029
- Klapwijk, E. T., Van De Kamp, F., Van Der Meulen, M., Peters, S., and Wierenga, L. M. (2019). Qoala-T: a supervised-learning tool for quality control of freesurfer segmented MRI data. *NeuroImage* 189, 116–129. doi: 10.1016/j.neuroimage.2019.01.014
- Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., et al. (2013). Human connectome project informatics: quality control, database services, and data visualization. *NeuroImage Map. Connectome* 80, 202–219. doi: 10.1016/j.neuroimage.2013.05.077
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro Resource for Sharing of Neuroscience Data. *eLife* 10, e71774. doi: 10.7554/eLife.71774.sa2
- Mortamet, B., Bernstein, M. A., Jack, C. R. Jr., Gunter, J. L., Ward, C., Britson, P. J., et al. (2009). Automatic quality assessment in structural brain magnetic resonance imaging. *Magn. Reson. Med.* 62, 365–372. doi: 10.1002/mrm.21992
- Power, J. D. (2017). A simple but useful way to assess fMRI scan qualities. *NeuroImage* 154, 150–158. doi: 10.1016/j.neuroimage.2016.08.009
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Provins, C., Alemán-Gómez, Y., Cleusix, M., Jenni, R., Richiardi, J., Hagmann, P., et al. (2022b). Defacing biases manual and automated quality assessments of structural MRI with MRIQC,” in *28th Annual Meeting of the Organization for Human Brain Mapping (OHBM)* (Glasgow). p. WTh566. doi: 10.31219/0sf.io/8mcyz
- Provins, C., Alemán-Gómez, Y., Richiardi, J., Poldrack, R. A., Hagmann, P., and Esteban, O. (2022c). “Defacing biases in manual and automatic quality assessments of structural MRI with MRIQC,” in *Peer Community in Registered Reports (Registered Report Under Consideration Toward Stage 1)*.
- Provins, C., Markiewicz, C. J., Ciric, R., Goncalves, M., Caballero-Gaudes, C., Poldrack, R. A., Hagmann, P., and Esteban, O. (2022a). Quality Control and nuisance regression of fMRI, looking out where signal should not be found. *Proc. Intl. Soc. Mag. Reson. Med.* 31, (ISMRM), pp. 2683. doi: 10.31219/0sf.io/hz52v
- Shehzad, Z., Givasis, S., Li, Q., Benhajali, Y., Yan, C., Yang, Z., et al. (2015). The preprocessed connectomes project quality assessment protocol-A resource for measuring the quality of MRI data. *Front. Neurosci. Conf. Neuroinformatics*. doi: 10.3389/conf.fnins.2015.91.00047
- White, T., Jansen, P. R., Muetzel, R. L., Sudre, G., El Marroun, H., Tiemeier, H., et al. (2018). Automated quality assessment of structural magnetic resonance images in children: comparison with visual inspection and surface-based reconstruction. *Hum. Brain Map.* 39, 1218–1231. doi: 10.1002/hbm.23911
- Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., van den Heuvel, M. P., and Breakspear, M., (2016). Connectome sensitivity or specificity: which is more important? *NeuroImage* 142, 407–420. doi: 10.1016/j.neuroimage.2016.06.035



OPEN ACCESS

EDITED BY

Xin Di,
New Jersey Institute of Technology,
United States

REVIEWED BY

Xiaoxiao Wang,
University of Science and Technology
of China, China
Xi-Nian Zuo,
Beijing Normal University, China

*CORRESPONDENCE

Richard C. Reynolds
✉ reynoldr@mail.nih.gov

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

RECEIVED 18 October 2022

ACCEPTED 16 December 2022

PUBLISHED 30 January 2023

CITATION

Reynolds RC, Taylor PA and Glen DR
(2023) Quality control practices
in FMRI analysis: Philosophy, methods
and examples using AFNI.
Front. Neurosci. 16:1073800.
doi: 10.3389/fnins.2022.1073800

COPYRIGHT

© 2023 Reynolds, Taylor and Glen. This
is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in
other forums is permitted, provided
the original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Quality control practices in FMRI analysis: Philosophy, methods and examples using AFNI

Richard C. Reynolds*, Paul A. Taylor and Daniel R. Glen

Scientific and Statistical Computing Core, NIMH, NIH, Bethesda, MD, United States

Quality control (QC) is a necessary, but often an under-appreciated, part of FMRI processing. Here we describe procedures for performing QC on acquired or publicly available FMRI datasets using the widely used AFNI software package. This work is part of the Research Topic, “Demonstrating Quality Control (QC) Procedures in fMRI.” We used a sequential, hierarchical approach that contained the following major stages: (1) GTKYD (getting to know your data, esp. its basic acquisition properties), (2) APQUANT (examining quantifiable measures, with thresholds), (3) APQUAL (viewing qualitative images, graphs, and other information in systematic HTML reports) and (4) GUI (checking features interactively with a graphical user interface); and for task data, and (5) STIM (checking stimulus event timing statistics). We describe how these are complementary and reinforce each other to help researchers stay close to their data. We processed and evaluated the provided, publicly available resting state data collections (7 groups, 139 total subjects) and task-based data collection (1 group, 30 subjects). As specified within the Topic guidelines, each subject’s dataset was placed into one of three categories: Include, exclude or uncertain. The main focus of this paper, however, is the detailed description of QC procedures: How to understand the contents of an FMRI dataset, to check its contents for appropriateness, to verify processing steps, and to examine potential quality issues. Scripts for the processing and analysis are freely available.

KEYWORDS

FMRI, quality control, AFNI, resting state, reproducibility, processing, data visualization, task-based

Introduction

Quality control (QC) is a vital part of FMRI analyses, although it is often not detailed in studies or presentations. The presence of poor quality data can reduce the power and generalizability of results. Undetected non-physiological artifacts can greatly skew outcomes and alter study results. Importantly, some exclusionary criteria could also systematically bias results away from an accurate interpretation of the data and underlying brain behavior.

In theory, FMRI QC appears to be a straightforward process: Sort a data collection into “good” datasets to use, and “bad” datasets to exclude. Some set of metrics or quantities can be calculated to do this screening automatically, and then processing can proceed with the good subset. In practice, however, QC is a notably more challenging procedure because of the combined complexities and varieties of both FMRI acquisition and analyses.

We consider QC to be an integral part of the processing itself, rather than a separate step, because what it means to be a “usable” dataset depends on the processing steps and design of the final analysis. Consider a few basic examples:

1. The cerebellum in a subject’s dataset is truncated by the acquisition field of view (FOV): This subject’s data might be included in the final analysis of a purely cortical study but excluded in the case of a cerebellar-specific or whole brain study.
2. EPI signal strength and distortions can vary across the brain. Having a low temporal signal-to-noise ratio (TSNR) within the basal forebrain region might exclude a subject from a subcortical study, but not from one of the visual cortex.
3. Subject motion is one of the most difficult effects to account for within any study, particularly in resting state FMRI where it can drastically influence results. How many time points can be censored before a subject is deemed to have “too much” motion to include, and does this number change if one is studying a group that is predisposed to motion (e.g., young adolescents or Parkinson’s disease patients)? And what is even the “correct” censoring limit to utilize?

In this paper we describe a number of QC measures for both task-based and non-task (e.g., resting state or naturalistic) FMRI processing that are implemented in the AFNI software suite (Cox, 1996). This paper is part of a community-wide FMRI open QC project, “Demonstrating Quality Control (QC) Procedures in fMRI,” where various groups of developers and researchers detail their own methods for QC of data. Specifically, we note the following goals and procedures from the Project description:¹

This project aims to showcase examples of QC practices across institutions and to foster discussions within the field. Here, we welcome researchers and developers across the globe to describe their QC methods in detail and to show them “in action” for a varied dataset acquired across multiple sites and scanners. . . We welcome researchers to

present their quality control assessments of the subjects in the provided data collection, listing which would be included or excluded from further analyses, and which might be considered borderline or “uncertain.”

Our own perspective is based on our individual and collective experiences as researchers, collaborators, educators and software developers of the AFNI toolbox. The design principle of the AFNI toolbox is, “To help keep researchers close to their data,” and this influences our view of QC measures, as well. Rather than viewing QC as simply filtering datasets into “good” or “bad” bins, we regard it as the larger procedure of *being as sure as possible about the contents of the data collection, from acquisition properties to artifact checking to regression evaluation*. We note that some QC steps are quantitative (they can be derived directly from one or more numbers), some are qualitative (e.g., they require visualization) or a combination. Some involve interactively investigating the datasets in a GUI, which can be facilitated in AFNI by scripting. Some QC items can be evaluated “per subject” and are essentially independent of any other member of the data collection, while others involve the relative comparison of a property.

Here, we detail a set of QC procedures for FMRI subjects and provide examples of applying these to the Project datasets. The first stage of QC can occur before any real “processing” of datasets has taken place, called “getting to know your data” (GTKYD). It is not necessarily part of inclusion/exclusion criteria, but it importantly ensures consistency of acquisition parameters and data properties. Next, systematic quantitative and qualitative stages are set up directly within afni_proc.py’s processing pipeline and QC HTML: APQUANT and APQUAL, respectively. For task-based FMRI, the STIM stage investigates the stimulus event and timing information. Finally, the GUI (graphical user interface) stage should always be used for some set of subjects in a study, to verify dataset properties in depth, and it can also be useful for investigating unknown features that may be found in other QC stages. In short, we implement a wide variety of QC procedures to be detailed, and we partition these into conceptual groupings in order to aid systematization. We aim to be as descriptive as possible, to provide a starter guide for possible QC during FMRI processing.

Methods: Data and processing

The datasets downloaded from the Project website and analyzed here were originally distributed as part of the following public repositories, according to the Project instructions: Functional Connectome Project (FCP; Biswal et al., 2010), ABIDE (Di Martino et al., 2014), and OpenNeuro (Markiewicz et al., 2021). They are due to be specifically identified in detail in a future publication of the Project, but we note that each subject’s dataset was acquired in a single session at 3T using a single echo

¹ See here for the main Project page: <https://www.frontiersin.org/research-topics/33922/demonstrating-quality-control-qc-procedures-in-fmri> and here for further details and download links for the datasets (<https://doi.org/10.17605/OSF.IO/QAESM>): <https://osf.io/qaesm/wiki/home/>.

EPI sequence, and overall they have fairly “typical” acquisition parameters (in terms of TR, voxel size, etc.—see below).

Here, AFNI v23.3.02 (Cox, 1996) and FreeSurfer v7 (Fischl and Dale, 2000) software packages were used for processing each of the resting state and task-based fMRI data collections. For each collection, AFNI’s `afni_proc.py` was used to set up the full fMRI processing pipeline, which runs through regression modeling and includes an automatically generated quality control (APQC) HTML report. The full set of scripts in each case are available online: https://github.com/afni/apaper_afniqc_frontiers.

As noted in the Introduction, some QC details rely on processing choices and on the analysis being performed. In the present Project, there was no stated group analysis, so we considered investigating these datasets in preparation for a generic cortical, voxelwise analysis. For this QC, we note issues regarding issues in cerebellum or midbrain, but do not exclude subjects based on these (these regions were often excluded or only partially included in the EPI FOVs).

Resting state fMRI data and processing

The provided resting state data collection consists of acquisitions from seven different sites, each of approximately 20 subjects, with a total of 139 subjects. Each site is signified by the hundreds-digit of the subject ID, by which we refer to each subset. That is, Group 1 contains sub-101, sub-102, etc.; Group 2 contains sub-201, sub-202, etc.

For each subject, there is one T1w anatomical and one EPI time series, except within Group 6, in which several (but not all) subjects have two EPI time series. The whole brain, T1w anatomical volumes typically have voxels with approximately 1.0 mm resolution, though there is some inter- and intra-group heterogeneity. The EPI time series have the following ranges of properties: TR = 2.0–2.5 s; minimum voxel edge = 1.56–4.00 mm, and maximum voxel edge = 3.10–4.00 mm (with varied anisotropy); in-plane matrix size = 64–128, and through-plane matrix size = 32–47; number of volumes (per run) = 123–724. Four out of seven sites had acquired (at least some) EPI and anatomical volumes obliquely. Further details about the heterogeneity of basic dataset properties are enumerated within the first stage of QC results (GTKYD), below.

The first step of processing was to run FreeSurfer’s `recon-all` on each T1w anatomical volume, providing an initial brain mask and parcellations for reference. FreeSurfer parcellations were entered into `afni_proc.py` as “follower” datasets, to be mapped to the final template space and to provide optional reference locations there. Note that if performing an ROI-based analysis, blurring would typically not be included in the processing steps. AFNI’s `@SSwarper` program was also run on each T1w volume, to provide both a final skullstripping (SS) mask and a non-linear warp [*via 3dQwarp*; Cox and Glen (2013)] from that anatomical

to the MNI-2009c (asymmetric) template space [Fonov et al. (2011)]. Identical `@SSwarper` commands were used for Groups 1–6, and for Group 7 a different cost function (nmi, normalized mutual information, instead of lpa, local pearson correlation absolute value) was utilized to improve results. These outputs of `@SSwarper` were included in the `afni_proc.py` command, described below.

AFNI’s `afni_proc.py` program was used to generate a full, reproducible fMRI processing pipeline across each Group. While the `afni_proc.py` command contains the specified “control variables” of each processing block, the created script (which is automatically commented) can also be read to understand the exact implementation details. Because each resting state group was acquired with slightly different parameters, particularly voxel size, individual `afni_proc.py` commands were created here for each so that parameters such as “applied blur” would be appropriate for each. In an expressly multisite study, which would combine subjects across all sites/Groups into a single analysis, this approach might differ—for example, one might apply an option to blur all EPI datasets to the same full-width at half-max (FWHM) value, for final uniformity. Here, the only parameters that varied across each group’s `afni_proc.py` commands were the values of the applied blur size (“-blur_size”) and final EPI voxel dimensions (“-final_dxyz”).

The `afni_proc.py` processing included initial despiking and slice timing correction. The EPI volume with the minimum fraction of outliers in the brain mask was selected to be a reference for motion correction (rigid-body alignment across the fMRI time series) and EPI-anatomical alignment (linear affine transformation with 12 degrees of freedom). EPI-anatomical alignment was calculated by first creating a brightness-homogenized version of the reference EPI volume and then using the “lpc+ZZ” cost function for local Pearson correlation (Saad et al., 2009). For anatomical-template alignment, the non-linear warp from the previous `@SSwarper` step was included. An EPI volume extents mask was applied to omit voxels that, due to motion, did not have acquired data throughout the entire time course. An EPI brain coverage mask was generated for the purpose of calculating statistics, but following the default behavior in `afni_proc.py`, this mask was not otherwise applied, leaving the time series basically unmasked, allowing for more complete QC (we recommend masking at the group level). A Gaussian blur was applied to each time series, with FWHM of approx. 1.5–2x the mean EPI voxel dimension (see scripts). Time series were scaled to have a mean of 100, to put the data in units of percent change. This scaling has a negligible effect on correlations, though it is helpful if computing parameters such as `fALFF`, for example.

The final processing block within the `afni_proc.py` command includes regression modeling, which amounts to projection of signals of non-interest, in the case of resting state analysis. This included censoring, for volumes with `Enorm`

(Euclidean norm of first differences of motion parameters) >0.2 mm or an outlier fraction $>5\%$ within a whole brain mask. Default polynomial regressors were used to model the slow baseline drifts. The six time series from rigid-body EPI alignment and each of their derivatives were included “per-run” as motion regressors. Bandpassing within the standard low frequency fluctuation (LFF) range of approx. 0.01 – 0.1 Hz was *not* included in this processing, since it has been shown that useful physiological data exist in the fMRI time series above 0.1 Hz (e.g., Gohel and Biswal, 2015; Shirer et al., 2015), and such bandpassing incurs a large statistical cost in terms of degrees of freedom (Caballero-Gaudes and Reynolds, 2017). The consequences for fMRI QC of including standard LFF range bandpassing are discussed below.

Task-based fMRI data and processing

The provided task-based data collection consists of 30 subjects (subject IDs: sub-001, sub-002, etc.) acquired at a single site. A single task paradigm was used, and timing files were provided in both original BIDS format and in a simplified, columnar format. For each subject, there is one T1w anatomical and one EPI time series. The whole brain, T1w anatomical volumes have 1.00 mm isotropic voxels. The EPI time series have the following properties: TR = 2.0 s; voxel dimensions = 3.00 mm \times 3.00 mm \times 4.00 mm; matrix dimensions = $64 \times 64 \times 34$; number of volumes = 242; oblique slices.

As for the resting state processing above, FreeSurfer recon-all and AFNI @SSwarper commands were run on each subject's T1w anatomical volumes. In setting up stimulus timing, we note that there are many ways to interpret and make a model from the event files. We chose to model the 2 event types, Task and Control, using reaction time for event duration, and the full duration if a subject did not respond in time. Control events had durations between 0 and 2 s, while task events lasted between 0 and 4 s. AFNI's timing_tool.py was used to apply this interpretation.

In the task-based afni_proc.py, the same processing blocks and options for slice timing correction, intra-EPI registration (for motion correction), EPI-anatomical alignment, anatomical-template alignment, mask estimation and scaling. The despiking block was not used. The blur size was set to 6 mm, the application of which was restricted to the estimated mask.

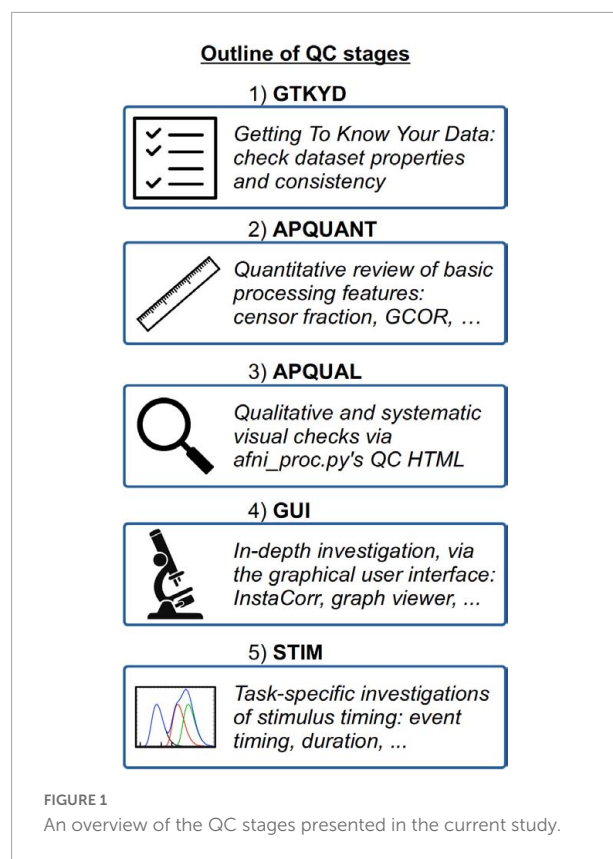
The regression model included censoring for volumes with $Enorm \geq 0.3$ mm (a slightly higher value than for the resting state processing, since the latter tends to be more sensitive to motion effects) or an outlier fraction $>5\%$ within a whole brain mask. The six time series from rigid-body EPI alignment were included per-run as motion regressors. In the task design, there were two stimulus classes: “Task” and “Control” events (the latter name should not be confused with the standard subject specification of “control group”; also, in

this Project, there were no such group classifications). These were modeled as duration modulated blocks, normalized to a 2 s response time [“-regress_basis_multi 'dmUBLOCK(-2)'”], and serial correlation within the time series was accounted for with 3dREMLfit (“-regress_reml_exec”). Two general linear tests (GLTs) were specified as potential conditions of interest: The contrast “Task - Control,” and the average stimulus response “ $0.5 \times (\text{Task} + \text{Control})$.”

General, simple and fast fMRI “quick” processing

The previous two sections describe the detailed processing options selected for the resting state and task-based processing commands implemented for these specific data collections. For each, several processing options and control parameters are selected by the user, tailored to the study design and research question. These are useful and appropriate for full dataset processing, e.g., as part of a group analysis.

However, we note an additional tool called ap_run_simple_rest.tcsh that is much simpler to set up for quick, general processing for any fMRI dataset; it is particularly useful for QC purposes. The AFNI program is a wrapper for afni_proc.py with a particularly simple front end: The only required options are the input dataset names



(some additional ones can be entered, too). Importantly, this program can be used to generate the vast majority of the QC information that is detailed below. In particular, almost every quantitative QC criterion (described under APQUANT) should be essentially identical.

This alternative analysis tool was designed with the focus of providing efficient checks for datasets as individual subjects are acquired, and can even be implemented to perform QC while the subject is in the scanner—thus, data could be reacquired if there were a particular problem such as severe motion or EPI dropout. Similarly, it could be easily created automatically as the scanner saves data to storage, to generate a uniform QC HTML report that would be immediately available to all researchers acquiring data. This tool uses affine template registration and processes data as resting state, making it simple, fast, and suitable to provide detailed QC. While the seed-based correlation QC maps can be considered slightly noisier than in a full processing case that implements non-linear alignment, they should still be reasonable and useful for quick QC purposes. In this work, we describe the QC items using the specific `afni_proc.py` commands, but the same considerations would apply to the “quick” outputs here.

Procedures for FMRI quality control

A schematic overview of the QC stages is shown in [Figure 1](#).

1) GTKYD: Getting to know your data

The first stage in the QC procedure here is referred to as GTKYD, which has two primary features. First, this checks the consistency of several key data and header properties within the group, such as dataset orientation, matrix size and more. Second, this investigates the reasonableness of the dataset properties, such as voxel size (units and isotropy), minimum/maximum values within the volumes (for possible scanner saturation) and more. Problematic values in either of these “relative” and “absolute” checks, respectively, might be a sign of acquisition mistake, DICOM-to-NIFTI conversion trouble, incorrect header information, BIDS construction, or other errors when creating the collection.

In general, this GTKYD stage is not intended to be used to include/exclude individual subjects. Instead, its purpose is to verify that the datasets contain their expected properties and are appropriate for the analysis at hand. Questions or potential issues should lead to double-checking the acquisition sequence and reconstruction steps, whether collected by the researchers performing the analysis, or, for public or shared data, by contacting those who did acquire it. In the first case, we recommend performing the QC steps immediately and

repeatedly as each subject in a study is collected, to protect against long-running and fundamental issues in the data, which may lead to wasted acquisition time and expense. In all cases, GTKYD reduces the possibility of analyzing fundamentally problematic or inappropriate data.

The GTKYD properties checked here included the following for both EPI and anatomical volumes:

- header-info: Matrix size, orientation, voxel dimensions, datum type, NIFTI `qform_code`, NIFTI `sform_code`
- data-info: Number of runs, minimum value, maximum value.

Additionally, the following was checked for EPI:

- header-info: TR, number of time points, slice timing.

2) APQUANT: Quantitative review of basic processing features

This stage describes the automation of quantitative QC measures output during `afni_proc.py` processing. This includes scriptable subject exclusion criteria, as well as checks for processing consistency and additional warnings. The output of this stage, created by AFNI's `gen_ss_review_table.py` (GSSRT) program, is a list of subjects to exclude/include.

During processing with `afni_proc.py`, a results directory is created for the full output, including storage of intermediate datasets, text files, and other information. In particular for this QC step, a single file of “basic” review quantities related to the processing is made. This essentially contains a dictionary of summary information about the processing—such as software version used, input datasets, censor fractions, and more—for each subject. For example, the “TRs censored” field records how many time points were censored during the subject's processing, “motion limit” records the threshold value used for Enorm censoring, and “global correlation GCOR” records the average correlation across all pairs of brain-masked voxels. These single subject review dictionaries can be combined across the group into an information table, using GSSRT, with one subject per row and one dictionary key (or review field) per column.

Importantly, one can provide a “checklist” of features to query, and create a sub-table of subjects that have one or more properties. For example, one could apply a set of exclusion criteria by generating a subtable of all subjects who have too many censored time points or too low an average TSNR. Additionally, one can create descriptive tables to verify that all subjects had similar EPI voxel sizes and were analyzed with the same software version. This combination of `afni_proc.py`'s basic processing dictionary and GSSRT's table-generating functionality is very flexible and useful for staying informed about a wide range of properties about the data processing as a whole.

Here, we created three separate review tables for each group: One for checking the analysis consistency across subjects; one for checking for possible concerns in the data at a “warning” level; and one for applying strict exclusion criteria. The GSSRT fields and comparison operators for each table’s checklist are shown in [Table 1](#). As noted in the table, all the same criteria were applied to both resting state and task-based fMRI collections, with one additional exclusion criterion for the latter. Additional criteria could be selected, as well, depending on the study. For example, while it was not used in this study, a Dice coefficient for the overlap between the EPI mask and the anatomical mask would be useful for cases where a specified minimum fraction of brain coverage is required.

It is important to note that the specific threshold values we have used for the quantitative keys could differ across studies. For example, rodent datasets would have much smaller head size and voxels, and one might expect less motion if they were anesthetized. One might allow for different motion criteria in a study of motion-prone children. The appropriateness of a particular TSNR threshold may vary with scanner. Over time, more knowledge may be accrued to inform better parameter selections, from the point of view of sensitivity and specificity. The present values seemed reasonable for this study and may form a starting basis for other ones, but should not be taken as absolute.

3) APQUAL: Qualitative and visual checks using afni_proc.py’s QC HTML

In complement to the APQUANT stage, this section describes performing a qualitative, visual-based assessment of the processing results. In particular, this is done using afni_proc.py’s QC report (APQC), which is an automatically generated HTML document. It is an interactive HTML for investigating various features of the data, including the original data, alignment, statistical maps and modeling, motion, warnings, and more. Ratings and comments can be saved for each QC block.

While some features of processing can be assessed quantitatively, many others essentially require visualization. For example, image registration is driven by a quantitative cost function, but then separate assessment is needed to verify that tissue boundaries and sulcal and gyral patterns appear to be well-aligned. Furthermore, there are numerous potentially artifactual patterns that can appear in datasets; these can be most easily identified by the human eye, and either recognized directly or marked for requiring further exploration. In many cases, fully understanding a subject’s dataset and problems that may exist with it requires having a multifaceted appreciation for it, and the APQC HTML provides one form of this.

The APQC HTML is organized in successive “QC blocks,” whose elements are grouped by processing steps and conceptual

relatedness. Most blocks are common to both task-based and non-task processing, though some features are distinct (as noted in the descriptions below). Additionally, some features depend on the details of processing—e.g., the anatomical-to-template alignment block only exists if one is registering the subject to a template space. In the following, we describe the current QC blocks and features for single-echo fMRI processing. For each block, we provide a list of elements or keys that describe specific features in a QC assessment, and these terms are used when evaluating the present data collection in the Results section. These keys may provide a generalizable categorization for QC reporting. They are also likely to grow in number over time.

vorig

*Views of the original space EPI (specifically, the volume registration reference) and anatomical volumes, as well as their overlap.**

- EPI: FOV coverage, signal dropout, ghosting overlap, poor tissue contrast (esp. if alignment fails), spatial distortion (see better check in ve2a), inhomogeneity.
- anat: FOV coverage, ringing, poor tissue contrast (esp. if alignment fails), inhomogeneity, skull stripping (if previously applied).
- overlap: Initial EPI/anat overlap (informational, in case EPI-anatomical alignment fails).

ve2a

*Views of the EPI-to-anatomical alignment results: Anatomical edges overlayed on the EPI.**

- global: Overall quality of alignment (e.g., from sulcal, gyral and ventricle patterns; note CSF can affect the appearance of the outer edge).
- local: Part of volume matching is poor (particularly around regions of interest), which can be due to FOV coverage, EPI signal dropout, distortion, other.

va2t

*Views of the anatomical-to-template results: Template edges overlayed on the anatomical.**

- global: Overall quality of alignment (e.g., from sulcal, gyral and ventricle patterns).
- local: Part of volume matching is poor due to, e.g., FOV coverage, distortion, SS, regional mismatch, other.

**One of the va2t, ve2a or vorig QC blocks will contain a view of the final EPI mask overlayed on the final reference volume, determined by whether the final space is a template, the subject’s anatomical or the subject’s EPI, respectively.*

TABLE 1 Lists of QC criteria for generating review tables of different properties after completing single subject processing with afni_proc.py.

APQUANT checklists (rest and task FMRI)		
Consistency checklist (rest, task)		
Key/field	Comp.	Description
'AFNI version'	VARY	Does the package version vary?
'num regs of interest'	VARY	Does the number of regressors of interest vary?
'final voxel resolution'	VARY	Do the final voxel dimensions vary?
'num TRs per run'	VARY	Does the number of EPI time points per run vary?
Warnings checklist (rest, task)		
Key/field	Comp.	Description
'final DF fraction'	LE 0.7	Is the remaining fraction of degrees of freedom ≥ 0.7 ? (NB: Bandpassing would affect this.) Visualize DF summary in APQC 'regr' block.
'censor fraction'	GE 0.15	Is the fraction of censored time points ≥ 0.15 ?
'average censored motion'	GE 0.1	After censoring, is the remaining average motion (Enorm) ≥ 0.1 mm?
'max censored displacement'	GE 6	Are any two volumes ≥ 6 mm apart?
'global correlation (GCOR)'	GE 0.15	Is GCOR ≥ 0.15 ? Visualize in APQC 'regr' block as corr_brain.
'TSNR average'	LT 150	Is the within-mask average TSNR ≤ 150 ? Visualize in APQC 'regr' block as TSNR-final.
Exclusion criteria checklist (rest)		
Key/field	Comp.	Description
'final DF fraction'	LE 0.6	Is the remaining fraction of degrees of freedom ≤ 0.6 ? (NB: Bandpassing would affect this.) Visualize DF summary in APQC 'regr' block.
'censor fraction'	GE 0.2	Is the fraction of censored time points ≥ 0.2 ?
'average censored motion'	GE 0.15	After censoring, is the remaining average motion (Enorm) ≥ 0.15 mm?
'max censored displacement'	GE 8	Are any two volumes ≥ 8 mm apart?
'global correlation (GCOR)'	GE 0.20	Is GCOR ≥ 0.20 ? Visualize in APQC 'regr' block as corr_brain.
'flip guess'	EQ DO_FLIP	Is there an EPI-anatomical left-right flip? Visualize in APQC 'warns' block.
Exclusion criteria checklist (task)		
Key/field	Comp.	Description
'final DF fraction'	LE 0.6	Is the remaining fraction of degrees of freedom ≤ 0.6 ? (NB: Bandpassing would affect this.) Visualize DF summary in APQC 'regr' block.
'censor fraction'	GE 0.2	Is the fraction of censored time points ≥ 0.2 ?
'average censored motion'	GE 0.15	After censoring, is the remaining average motion (Enorm) ≥ 0.15 mm?
'max censored displacement'	GE 8	Are any two volumes ≥ 8 mm apart?
'global correlation (GCOR)'	GE 0.20	Is GCOR ≥ 0.20 ? Visualize in APQC 'regr' block as corr_brain.
'flip guess'	EQ DO_FLIP	Is there an EPI-anatomical left-right flip? Visualize in APQC 'warns' block.
'fraction TRs censored'	GE 0.2	Is the fraction of time censored from any stimulus response ≥ 0.2 ?

These key or field values are automatically placed in a text file within each subject's results directory by afni_proc.py. Each set of key/fields and comparisons (Comps.) is then used within a gen_ss_review_table.py command to create a summary table. The following comparison operators are used here: VARY = "differs across subjects"; GE = "greater than or equal to"; LE = "less than or equal"; and EQ = "equal to." For each group of subjects, a set of consistency, warning and drop criteria tables were made.

- mask-overlap: Estimated coverage of usable FMRI signal (typically intersected with the subject anatomical).

vstat

Views of relevant statistical modeling. For non-task FMRI, when a recognized template space is used, seed-based correlation maps of the default mode, visual and auditory networks are shown. For task-based FMRI, the views include the full F-stat of modeling, as well as coefficient + stat maps of stimuli and contrasts of interest.²

- quality: Overall expected/recognizable network correlation (or task statistical) patterns observed, such as full spatial coverage (no missing regions); network specificity (no extra regions); reasonable magnitude; extra-cranial patterns.
- artifact: Ghosting; striping; strong slice-based patterns; large spatial patterns across/unconstrained by tissue type; notably non-physiological patterns.

We note that in these images, and in several others within the APQC HTML, the thresholds are applied transparently. That is, suprathreshold regions are shown opaque (or with maximum translucency) and outlined, and subthreshold values are displayed with increasing transparency as the magnitude of the value decreases. This reduces the sensitivity to choice of threshold, and allows focal regions to be highlighted (with opacity and outlining) while still showing information throughout the brain (Allen et al., 2012; Taylor et al., 2022). Moreover, brain masks are typically not applied, to show results throughout the full FOV, which helps to further identify any potential artifacts.

mot

Motion-related information: Plots of Enorm, outlier fraction, and motion parameter time series (with any censoring information shown), and a grayplot of residuals. Provides a useful reference (censor- and motion-related quantities are primarily checked across the group using GSSRT).

- enorm: Odd patterns; regular signals, which are likely not physiological (e.g., mechanically driven); many time points with just sub-threshold values, which might drive spurious correlation (might lead to re-processing); overall value range.
- outliers: (same items as “enorm,” above); evaluate for synchrony against enorm.
- volreg-pars: Similar properties to “enorm” above; note that these parameters are not directly thresholded for censoring.

- grayplot: Strong vertical patterns may suggest high residual correlation (primarily checked in “regr-corr_brain” visualization and quantified in “qsumm” with GCOR).

regr

Regression modeling information: Degree of freedom (DF) summary; view of correlation map with whole brain average residual signal (checks brainwide similarity of residuals, such as for large breathing, and motion effects remaining); and TSNR maps (good scenario: Relatively consistent TSNR around brain regions of interest). For task datasets, plots of the individual regressors of interest, as well as their sum, are shown (with any censor bands, for reference).

- task-ideal-sum: (NB: Strongly paradigm dependent) any problem with the sum of regressors; large gaps and/or spikes might generally be worth noting.
- task-ideal-stim: (NB: Strongly paradigm dependent) any problem with an individual regressor of interest; duplicated stim timing (scripting mistake); stimulus-correlated motion may be worth noting.
- df-count: Too few degrees of freedom in output results (often due to censor fraction and/or bandpassing); typically checked automatically with GSSRT.
- corr_brain-artifact: (similar to vstat-artifacts) ghosting, correlation/anticorrelation striping, strong slice-based patterns, large spatial patterns across/unconstrained by tissue type, notably non-physiological patterns.
- corr_brain-quality: Too high (also typically quantified via GCOR and checked with GSSRT) or too low.
- TSNR_volreg: Mainly informational, since this is calculated before regression modeling and noise regression (look for similar features as in TSNR_final-* items).
- TSNR_final-loss: Notable dropout/low signal in regions of interest (e.g., often low in frontal/temporal lobes and subcortical nuclei).
- TSNR_final-artifact: Non-physiological patterns of TSNR magnitude, particularly dropout (e.g., vertical bands/stripes).

radcor

Radial correlation maps: The value of correlating each voxel with a Gaussian-weighted local average (FWHM = 40 mm in human datasets). A typically good scenario is relatively high values approximately constrained to GM; motion effects often appear as high correlation/anticorrelation patterns around the edge of the brain, which are often reduced after volreg.

- tcat-artifact: Mainly informational, since this is calculated for initial data with no motion correction (look for similar features as in radcor_volreg-* items).

² By default, afni_proc.py creates images of the full F-stat and up to 4 additional coefficient + statistics pairs, depending on the number of stimuli and contrasts in the regression model. The user may specify any number of stimulus and contrast results to show, however.

- **volreg-artifact:** Patches of high radcor values spanning multiple tissue types (can be sign of coil artifact or other non-physiological effects); artifacts here often inspire investigations with InstaCorr, as referred to in Procedure 4.

warns

List of warnings from various checks throughout processing, including for: Regression matrix correlations; high censor fractions; pre-steady state outliers; left-right flip between input EPI and anatomical. Several can be checked with GSSRT (with useful details here for verification).

- **regr_mat:** High pairwise correlations in regression matrix (varied).
- **gen-censor:** High total/overall censoring fraction (typically checked with GSSRT).
- **task-stim-censor:** High censoring fraction for one or more particular stimuli (optionally checked with GSSRT).
- **press:** EPI data appear to have pre-steady state volumes at the start (*via* outlier check; though sometimes this is simply due to motion in first time points).
- **flip:** EPI-anatomical might have relative flip, as checked with cost function alignment and to-be verified with provided images (Glen et al., 2020).

qsumm

Basic quantitative information of processing, such as AFNI software version, voxel sizes; motion limits and counts; TSNR; and more. Provides a quick reference (many of these quantities should be checked across the group using GSSRT).

- **anomalous:** An unexpected value, such as final voxel resolution, software version number, etc.
- **suprathresh:** Unexpectedly or problematically large value (e.g., censor fraction, GCOR).
- **subthresh:** Unexpectedly or problematically small value (e.g., average TSNR, maximum F-stat).
- **missing:** Quantity not present, perhaps due to coding error (e.g., missing censor fraction, missing censor fraction per run).

4) GUI: In-depth investigation with the graphical user interface

This stage describes exploring one or more datasets interactively. While this may require more time to perform than some other steps, it provides the best means for understanding things like the detailed alignment of two volumes, the combined spatio-temporal aspects of EPI time series (with “InstaCorr,” described here), etc. To facilitate this process, `afni_proc.py` automatically generates multiple scripts to load particular datasets and visualization functionality in the AFNI GUI.

- **align:** Check alignment (or registration) features.
- **graph:** View the time series plots of one or more voxels.
- **instacorr:** Flag peculiar spatio-temporal patterns in the time series data.
- **other:** Any other feature(s) using the afni and/or suma GUIs, plugins, etc.

align

There are a large number of features in the AFNI toolbox and GUI to inspect the alignment or registration between two datasets (e.g., see Appendix A in the [Supplementary material](#) of Glen et al., 2020). The default method is to show one volume as a grayscale background (underlay) dataset, while the other is shown in color as the “overlay” dataset. There are several methods for viewing the datasets interactive in different ways, depending on the properties of the datasets (matching or differing tissue contrasts, blurriness, etc.), which can help to focus on various features. These include: Toggling the underlay/overlay datasets, adjusting underlay contrast/brightness, adjusting overlay opacity, viewing the underlay edges, using a horizontal or vertical “image comparison” slider bar, and using a slider to fractionally blend the datasets.

graph

The AFNI GUI includes an interactive and expandable graph window for displaying the time series of one or more voxels. Observing properties of the time series, even when no stimulus has been provided, can provide useful insight, particularly into possible artifacts or non-neuronal confounds. For example, subject motion effects can be observed as peaks and sudden shifts in the amplitudes across many voxels. Drift or shimming-related changes can also be noted. One can also load a reference time series (e.g., one with the ideal task response) and investigate patterns, similarity or possible features showing stimulus-correlated motion.

InstaCorr

The InstaCorr functionality (which stands for “Instant Correlation”) within the AFNI GUI is the prime tool for an in-depth investigation of a 4D EPI dataset. Briefly, InstaCorr allows one to freely explore spatio-temporal patterns within a dataset by clicking and dragging a seed location anywhere throughout the volume; the resulting seed-based correlation patterns update continually and instantaneously, so that one can quickly assess a full FOV (see, e.g., Jo et al., 2010; Song et al., 2017). This is particularly useful for exploring functional networks, potential scanner artifacts, and more. Processing features such as baseline regression, bandpassing, smoothing, masking and setting a seed radius can all be selected within the InstaCorr setup menu. The `afni_proc.py` processing now automatically creates a “`run_instacorr*.tclsh`” script to run InstaCorr on the

regression model's output; running the script automatically opens the AFNI GUI with InstaCorr setup on the residuals dataset.

Here, we used InstaCorr in conjunction with the APQUAL step, when one or more QC images showed a questionable pattern. For example, we could observe whether there were: Large, non-physiological patches of high correlation; slice-constrained artifacts; and more. Resting state fMRI analysis often depends on correlation patterns, making InstaCorr verification particularly important. In task-based fMRI, it can provide useful exploration of areas where responses are unexpectedly low.

5) STIM: Task-specific investigations of stimulus timing

This stage describes understanding and evaluating the stimulus event timing for a task-based analysis. This includes answering whether events are presented at consistent intervals or randomized, of consistent duration or variable, and based on the subjects or not, both for duration and possibly amplitude modulators in the regression model. It includes answering similar questions for inter-stimulus intervals (ISIs). And it includes evaluating the stability of the regression matrix, i.e., whether small noise fluctuations could have a noticeable effect on the results.

There are several tools within AFNI that can be helpful for investigating various stimulus related features across the group, such as summaries of timing, duration and interstimulus intervals. These can be particularly useful in understanding variations or potential issues in subject results. Such investigations are essential during an experiment design phase, before acquiring subject data, and are similarly important for understanding event timing in a study from an external group, or even in review. Detailed investigations can be done for just a few subjects, while statistical reviews of stimulus durations and interstimulus interval timing can be performed and then summarized across all subjects, while looking for peculiarities or outlier subjects.

Two items that are often computed after the regression matrices exist are regressor correlations and condition numbers. Negative pairwise correlations are often expected, particularly in cases with two or just a few stimulus classes. As a measure of predictability, this happens when one stimulus response is “on” and another stimulus response is generally “off,” or lower. Such a pair of regressors might have a modestly high, negative correlation that is considered acceptable. Condition numbers (of the full model and conceptual sub-models) help identify when a model is becoming mathematically unstable, often from a stimulus design mistake, or by having too little non-stimulus time.

- events: (for just a few subjects) visually review event timing across all classes together, including onsets times, durations, and offsets from previous events, along with any modulators.
- stim-stats: Show min/mean/max/stddev of stimulus durations, per class and subject.
- isi-stats: Show min/mean/max/stddev of interstimulus intervals, per subject.
- X-cormat: (done in APQUANT.warns section, above) look for large pairwise correlations among the regression matrix regressors.
- X-cond: Look for high condition numbers across subsets of the regression matrix, including the baseline, motion terms, regressors of interest, and combinations of these sets up to the full matrix.

Results for resting state data collections

GTKYD summary

GTKYD was the first stage of checking each group's data. In the present study, no subjects were excluded because of this stage's results, but they did inform some processing choices (and in other cases, they indeed might lead to a group not being included in a study). **Table 2** shows a summary of basic dataset properties that were inconsistent across a group. For example, in Group 5 six out of 20 subjects have an anatomical volume with differing orientation. This may reflect acquisition or reconstruction inconsistency, but importantly it may hide an error in correctly assigning directionality within the volume. While most mistaken “flips” of directionality within a dataset can be quickly detected visually, this is not so for left-right flips; for humans, relative EPI-anatomical flips can typically be reliably detected (Glen et al., 2020), but this is not the case for animal datasets or when all datasets for a subject are flipped.

Surprisingly, most groups (5 out of 7) contain heterogeneity of at least one basic dataset property. In Group 5, the EPI voxel dimensions of five subjects differ notably, which will affect SNR throughout the brain; additionally, the high anisotropy of the five outlier subjects can produce artifacts due to alignment and regridding. In Group 6, the numbers and lengths of runs vary within the group in complicated ways. These forms of heterogeneity can affect the statistical properties of estimated quantities, and lead one to question the appropriateness of combining these subjects in a group analysis (when not performing an explicitly large, multisite study, and these differences have a larger relative variance within the paradigm). Each of these items should lead to checking with the source of the data. If acquiring the data locally, performing the GTKYD check with each new subject can help identify problems or changes immediately, and minimize data waste.

TABLE 2 Summary of the first stage of resting state FMRI QC: GTKYD ("getting to know your data").

GTKYD: "Getting To Know Your Data" results (resting state FMRI)	
Property	Description
Group 1: EPI	
matrix size diff	sub-118 has 112×112×47, from group std 96×96×47
num vols diff	sub-114 and sub-115 have 128, from group std 256
vox dim diff	sub-118 has 2.29×2.29×3.0 mm ³ , from group std 2.67×2.67×3.0 mm ³
Group 1: anatomical	
matrix size diff	sub-104, sub-109, sub-112 and sub-117 have 256×180×256, from group std 256×200×256
Group 2: EPI	
large max values	approx. 2–4×10 ⁶
oblique	
Group 2: anatomical	
vox dim diff	sub-203 has 1×0.93×0.93 mm**3 from group std 1×1×1 mm**3
matrix size diff	sub-118 has 160×288×288, from group std 160×256×256
Group 3:	
no warnings	
Group 4: EPI	
no slice timing	
Group 5: EPI	
matrix size diff	sub-501, sub-502, sub-503, sub-504, sub-509 have 128×128×34, instead of group norm 80×80×35; others have 80×80×35 and 80×80×39 mm ³ instead of group std 3.0×3.0×4.0 mm ³
datum diff	sub-501, sub-502, sub-503, sub-504 and sub-509 have float, instead of group std short
(some) oblique	
no slice timing	
Group 5: anatomical	
orient diff	sub-501, sub-502, sub-503, sub-504, sub-509 and sub-519 have RPI, instead of group std LPI
matrix size diff	much heterogeneity
oblique	
Group 6: EPI	
diff num of EPI	sub-601, sub-602, sub-603, sub-604, sub-605, sub-606, sub-607 and sub-620 only have 1, instead of group std 2
diff length of EPI	sub-601, sub-602, sub-603, sub-604, sub-605, sub-606, sub-607 and sub-620 have 240, 360, 480 or 724 time points, instead of group standard 130-133
oblique	
no slice timing	
Group 6: anatomical	
matrix size diff	sub-601, sub-602, sub-603, sub-604, sub-605, sub-606, sub-607, sub-612, sub-619, and sub-620 have 256×256×256, split with others having 256×256×176
oblique	

(Continued)

TABLE 2 (Continued)

GTKYD: “Getting To Know Your Data” results (resting state fMRI)	
Group 7: EPI	
oblique	
Group 7: anatomical	
(some) oblique	

For each group, this displays cases of heterogeneity in basic dataset properties, as well as noteworthy values for checking or for informing processing choices. Items shown here might prompt verification with the source of the data collection, whether it has been downloaded from a shared repository or is being acquired locally.

Table 2 also contains absolute quantities that were notable either to prompt verification from the source of the data or to inform processing choices. As an example of the former, Group 2’s EPI values ranged from zero to over 2×10^6 ; while fMRI datasets have no inherent units and this may not be a problem, these values are three orders of magnitude larger than typical dataset values, and therefore worth verifying their acquisition and reconstruction parameters to ensure that no numerical features (truncation, saturation, loss of contrast) have been introduced. Additionally, the EPI datasets in Groups 4, 5 and 6 did not contain slice timing information, which can be used for minor adjustment across the slicewise acquisitions. The lack of this information may be a reconstruction or distribution oversight, and hence obtainable. Finally, different software packages utilize obliquity information (the coordinate information that describes whether a dataset is acquired obliquely, away from simple cardinal orientations) differently during processing, such as: Applying it and regridding the data; ignoring it and effectively shifting coordinates; or leaving it in the header to be applied. Therefore, the choice and order of processing steps, particularly when it is present in an anatomical volume, may be affected by its presence. Here, we chose to remove obliquity of any anatomical volumes (while preserving the coordinate origin) as an initial processing step, to avoid issues with other software.

APQUANT evaluation

The quantitative drop criteria listed in Table 1 were applied to the processed data, followed by APQUAL evaluations for each subject and, in several cases, GUI checks. A brief summary table of applying these stages of QC to the afni_proc.py-processed datasets is shown in Table 3, listing subjects in one of the three specified categories: Include (“high confidence to use in the hypothetical study”), Exclude (“high confidence to remove”) and Uncertain (“there is a question about whether to include”). The Supplementary Table 1 contains a table with more detailed descriptions for each subject.

In these tables, the QC comments are named hierarchically, in the following format: STAGE.type[subtype](detail), using the terms listed in the previous section. For example,

APQUANT.excl.(“flip guess”) represents the label for the left-right flip check within the exclusion criterion check during the APQUANT stage. Some “detail” elements are not contained within the brief table, but are included in the more complete Supplementary Table 1. This notation has been introduced to provide a clear, brief reference to the source of the particular QC criterion.

There were 139 total resting state subjects processed. As discussed further below, Groups 2 (20 subjects) and 4 (23 subjects) were found to have artifacts across all subjects, following APQUAL and GUI QC checks. Of the remaining 96 subjects, 42 were categorized to include in further analysis, 37 to exclude, and 17 were listed as uncertain. Of the 37 to exclude, 31 were evaluated as such using APQUANT criteria: 21 by censor fraction, 8 by GCOR, and 2 by left-right flip checking (though one additional subject was categorized as “uncertain,” primarily due to left-right flip checking, as discussed in the APQUAL section below). The left-right flip evaluations were always visually verified during the APQUAL stage. The quantitative GCOR value typically correlates highly with the APQUAL’s “regr.corr_brain” evaluation, as well.

The warning-level APQUANT criteria were additionally noted in subject evaluations (see the detailed Supplementary Table 1). In particular, these were combined with APQUAL criteria for determining additional “exclusion” or other categorizations, as described below.

APQUAL evaluation

Figures 2–10 contain example images of the APQUAL evaluations, which are (by definition) qualitative and visual. Each figure shows multiple examples of the same APQC block from the HTML report. Each QC image is labeled with a colorband along its side, based on whether it would lead to excluding the subject (red), including the subject (green) or uncertain evaluation (yellow). Many images also contain arrows highlighting features of note.

vorig

Figure 2 shows QC examples from looking at one volume of the original EPI data (here, the “minimum outlier” volume from the EPI time series, which had the fewest outliers within the

TABLE 3 A brief summary of resting state fMRI dataset evaluations, based on the APQUANT, APQUAL and GUI QC checks.

QC evaluations (brief): Groups 1-7 (resting state fMRI)					
Group 1 (I = 7, E = 8, U = 5)			508	E	APQUAL.vstat.artifact
<i>sub</i>	<i>eval</i>	<i>comment</i>	509	E	APQUAL.vorig.EPI
101	E	APQUANT.excl('flip guess')	510	I	
102	U	GUI.instacorr(odd patterns)	511	E	APQUANT.excl('censor fraction')
103	I		512	E	APQUANT.excl('censor fraction')
104	E	APQUANT.excl('censor fraction')	513	U	APQUAL.vorig.EPI
105	I		514	I	
106	E	APQUANT.excl('censor fraction')	515	I	
107	U	APQUAL.vorig.EPI(ringing feature)	516	I	
108	I		517	U	APQUAL.vorig.EPI
109	I		518	E	APQUAL.vorig.EPI
110	U	APQUAL.vstat.quality	519	E	APQUAL.vorig.EPI
111	E	APQUANT.excl('GCOR')	520	I	APQUAL.regr.tsnr_final.quality
112	I				
113	I		Group 6 (I = 10, E = 7, U = 3)		
114	E	APQUAL.vstat.artifact	<i>sub</i>	<i>eval</i>	<i>comment</i>
115	E	APQUANT.excl('flip guess')	601	E	APQUANT.excl('GCOR')
116	E	APQUAL.warn.flip	602	I	
117	U	APQUAL.regr.TSNR_final-artifact	603	E	APQUANT.excl('GCOR')
118	E	APQUANT.excl('censor fraction')	604	I	
119	I		605	I	
120	U	APQUAL.regr.TSNR_final-artifact	606	E	APQUAL.regr.corr_brain-quality
Group 2 (I = 0, E = 20, U = 0)			607	I	
<i>sub</i>	<i>eval</i>	<i>comment</i>	608	I	
2*	E	GUI.instacorr('scanner artifact?')	609	E	APQUANT.excl('GCOR')
Group 3 (I = 9, E = 5, U = 2)			610	E	APQUANT.excl('GCOR')
<i>sub</i>	<i>eval</i>	<i>comment</i>	611	I	
301	U	APQUAL.vstat.quality	612	E	APQUANT.excl('GCOR')
302	I		613	E	APQUANT.excl('GCOR')
303	I		614	I	
304	I		615	U	APQUAL.regr.corr_brain-quality
305	U	APQUAL.vstat.quality	616	I	
306	I		617	I	
307	E	APQUANT.excl('censor fraction')	618	U	APQUANT.warn('GCOR')
308	I		619	U	APQUAL.vstat.quality
309	E	APQUANT.excl('censor fraction')	620	I	
310	I		Group 7 (I = 9, E = 10, U = 1)		
311	I		<i>sub</i>	<i>eval</i>	<i>comment</i>
312	I		701	E	APQUANT.excl('censor fraction')
313	I		702	I	
314	E	APQUANT.excl('censor fraction')	703	E	APQUANT.excl('censor fraction')
			704	U	APQUAL.vstat.quality

(Continued)

TABLE 3 (Continued)

315	E	APQUANT.excl('censor fraction')	705	E	APQUANT.excl('censor fraction')
316	E	APQUANT.excl('censor fraction')	706	E	APQUANT.excl('censor fraction')
			707	I	
Group 4 (I = 0, E = 23, U = 0)			708	E	APQUANT.excl('censor fraction')
sub	eval	comment	709	I	
4*	E	GUI.instacorr('scanner artifact?')	710	I	
			711	I	
Group 5 (I = 7, E = 7, U = 6)			712	E	APQUANT.excl('censor fraction')
sub	eval	comment	713	E	APQUANT.excl('censor fraction')
501	U	APQUAL.regr.TSNR_final-artifact	714	E	APQUANT.excl('censor fraction')
502	U	APQUAL.regr.TSNR_final-artifact	715	E	APQUANT.excl('censor fraction')
503	U	APQUAL.vstat.quality	716	E	APQUANT.excl('censor fraction')
504	U	APQUAL.regr.TSNR_final-artifact	717	I	
505	I		718	I	
506	I		719	I	
507	E	APQUANT.excl('censor fraction')	720	I	

The following abbreviations for evaluations ("eval") are used: E, exclude; I, include; U, uncertain. Both Groups 2 and 4 were found to have artifacts in each of their datasets, and hence all categorized for exclusion. A more detailed summary is provided in the [Supplementary Table 1](#), with further comments about most subjects.

brain mask and was used as a reference for motion correction and alignment to the anatomical). In panel A, sub-315's EPI shows a medium-sized patch of signal dropout. The associated anatomical volume contained a smaller spot at that location, so it is likely due to some local object (rather than a scanner artifact). This places a question of the full signal effects in this region, but since it is only moderate size and relatively constrained to the central sulcus, it might be reasonable to include the subject.

In [Figure 2B](#), there is a strong ghosting signal present, as further investigated using InstaCorr. It is particularly noticeable throughout the central region of the brain, and, therefore, the signal patterns would be highly non-physiological, and the subject should be excluded. The subject in panel C has a smaller amount of ghosting and a "ringing" artifact in the inferior slices. The exact degree of signal effect is uncertain, hence the QC rating. In panel D, we see that sub-509 has extremely large ventricles, which reduce the quality of anatomical-to-template alignment, and may also reduce the quality of EPI signal. The subject also has a large amount of frontal and subcortical signal dropout, which renders inclusion uncertain.

Finally, there were multiple subjects in Group 5 who had upside-down EPI volumes, as shown in [Figure 2E](#). Such large header errors warrant rejection, because the correct left-right designation is not possible to reliably ascertain *a posteriori*, without a marker. While it would be possible to try to fix the header and then assess results against the subject's anatomical using AFNI's left-right flip check, given the nature of this header

issue one might not be sure of the correctness of the anatomical volume's reconstruction. Therefore, given the high uncertainty of basic properties, such subjects should likely be excluded (though, in a different setting, one might contact the source of the data and query whether the initial reconstruction could be corrected).

ve2a

[Figure 3](#) shows the alignment of an EPI volume (underlay) to the same subject's anatomical (overlay, as edges). While EPIs typically contain geometric distortions (e.g., EPI distortion along the phase encode axis), affine registration is typically adequate to align most major structures to the higher-resolution and -detailed anatomical, as shown in panel A. However, EPI images often contain signal dropout, particularly bordering the sinus cavities, bordering the orbitofrontal cortex and subcortex. The ve2a block (views of EPI-anatomical alignment) provide useful images for assessing locations of dropout (as do TSNR maps, described below). Panel B shows several locations of poor signal strength and attenuation, which renders the suitability of sub-210's data uncertain. Panel C shows a case where the geometric distortions make global EPI-anatomical alignment difficult (see the signal pileup in the anterior and attenuation/extension in the visual cortex).

An important point for judging EPI-anatomical alignment is exemplified in [Figure 3D](#). The most important features to verify as matching are the tissue boundaries, sulci and gyri: The internal structures. At the edge of the brain, cerebrospinal fluid (CSF) can variously appear brightly, and make alignment details

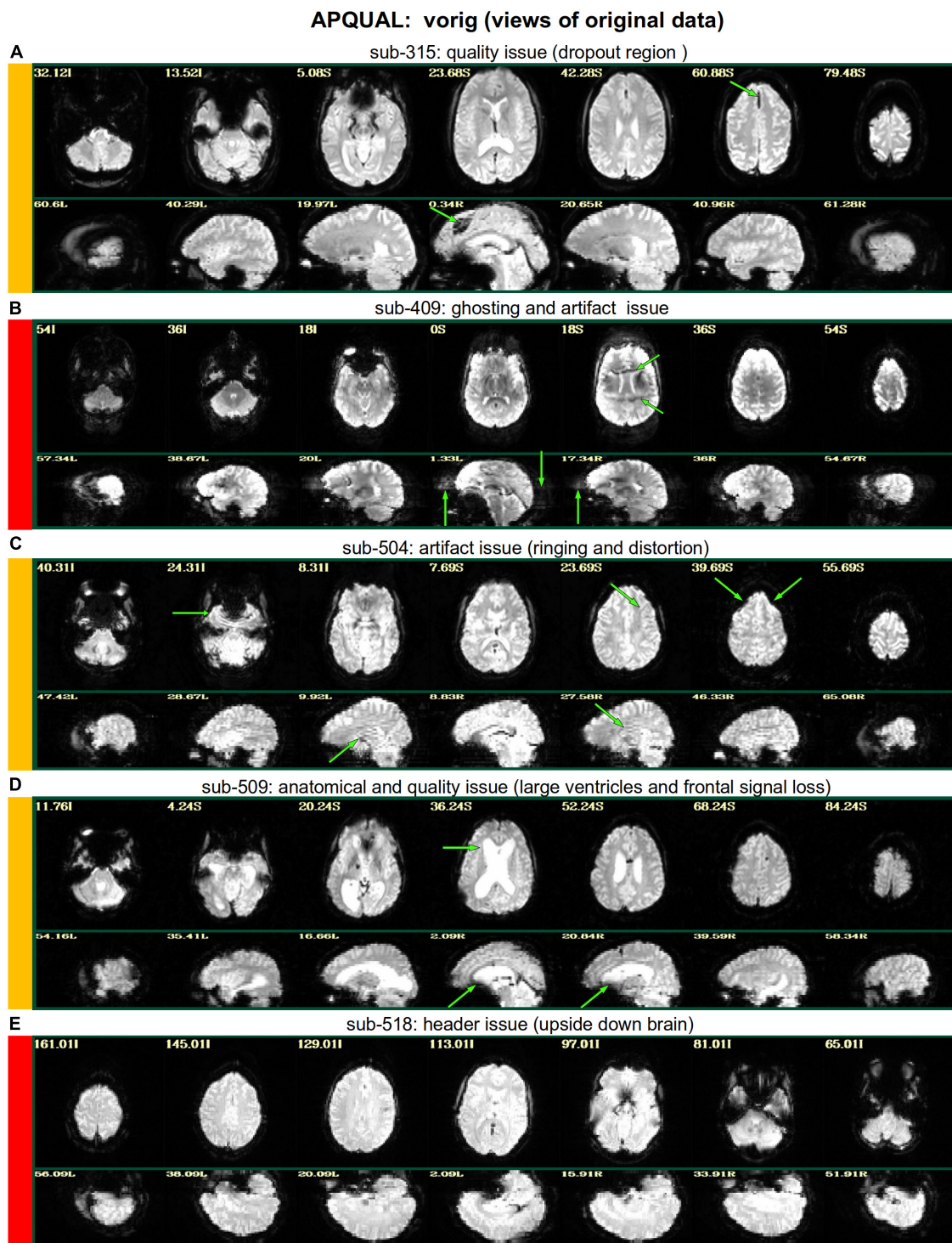


FIGURE 2

APQUAL examples for the “vorig” QC block: Visualizations of the original datasets (here, just the EPIs). In this figure and below, the colored bands to the left of each item denote whether the given QC item would suggest that the subject should be excluded (red), included (green) or leads to an “uncertain” evaluation (yellow); also, see [Table 3](#) for brief, overall evaluations for each subject, and the [Supplementary Table 1](#) for detailed QC comments. (A) The EPI contains a moderately sized dropout region (but it is mostly contained within the central sulcus). (B) This EPI contains severe ghosting artifact. (C) The inferior slices show a ringing artifact, and the frontal region is geometrically distorted. (D) This subject’s large ventricle may negatively affect alignment to template space, and there is notable dropout in the orbitofrontal region and subcortex. (E) The EPI is upside down, a significant header or data conversion problem.

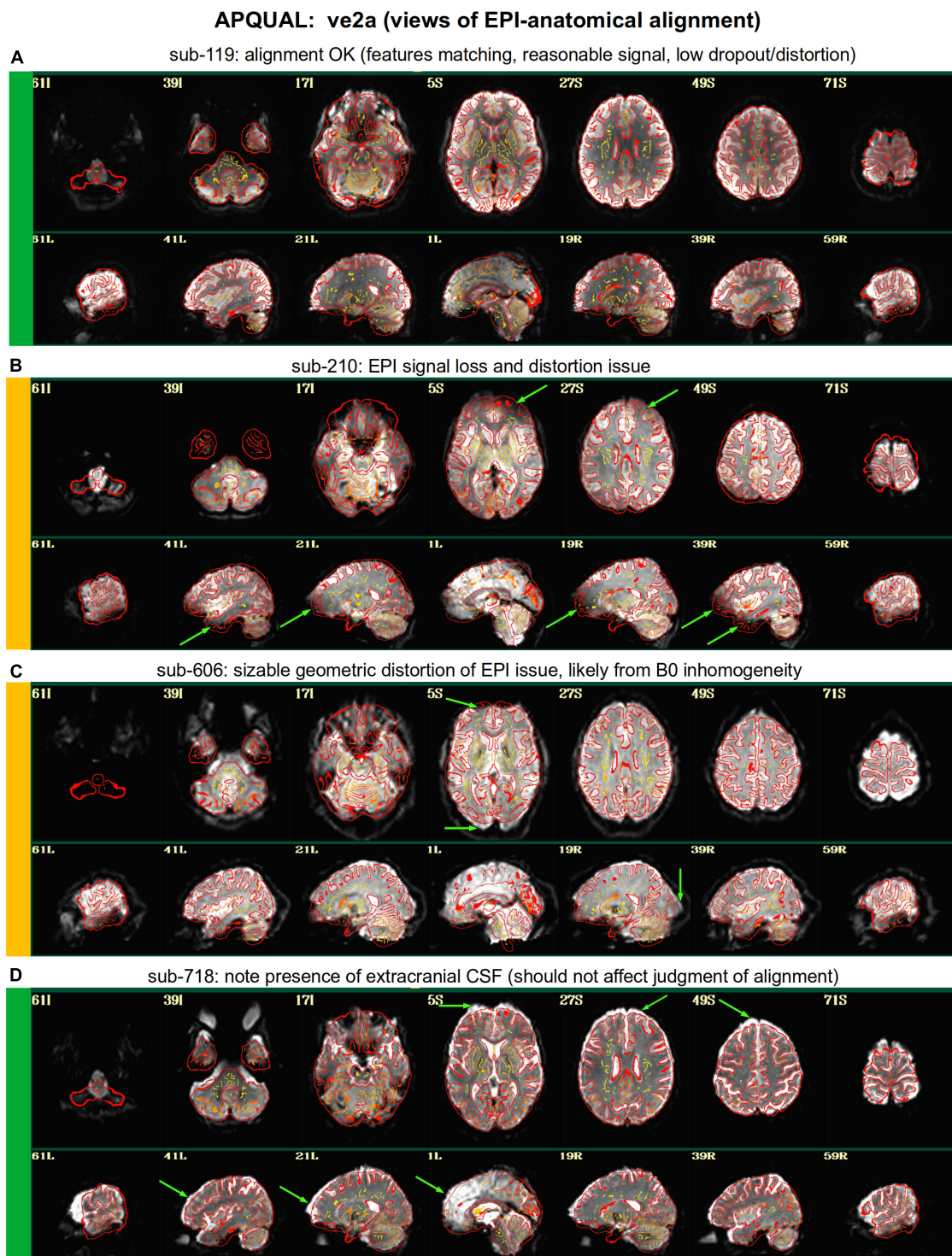
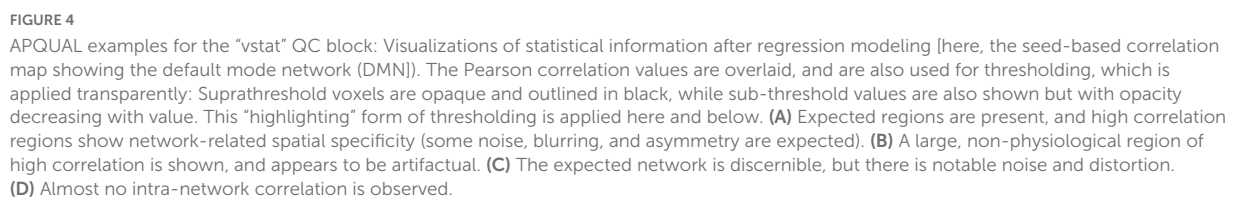


FIGURE 3

APQUAL examples for the “ve2a” QC block: Visualizations of the EPI-to-anatomical alignment (underlay = EPI; overlay = anatomical edges). (A) Structures appear generally well-registered. (B) There is notable EPI signal loss in the frontal and subcortical regions. (C) The EPI contains large distortions: Signal pileup in the anterior, and geometric stretching and signal attenuation in the visual cortex. (D) In judging EPI-anatomical alignment, interior structures matter most and CSF (bright, and highlighted with arrows) should be ignored.



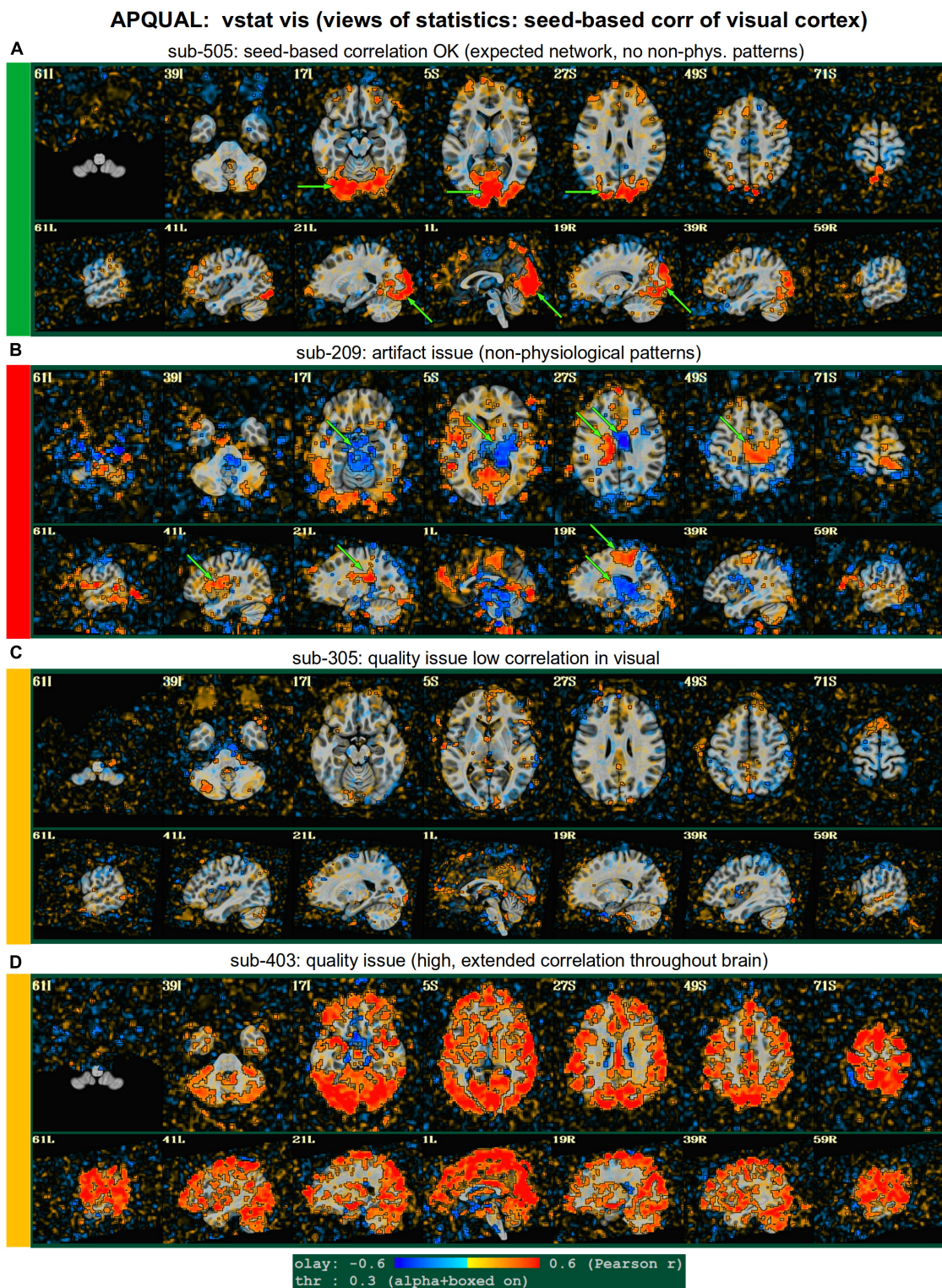


FIGURE 5

APQUAL examples for the “vstat” QC block: Visualizations of the statistical information after regression modeling (here, the seed-based correlation map showing the visual network). (A) Expected regions are present, and high correlation regions show network-related spatial specificity (some noise, blurring, and asymmetry are expected). (B) A large, non-physiological region of high correlation is shown, expanding across multiple tissue boundaries, and appears to be artifactual. (C) Almost no intra-network correlation is observed. (D) The high correlation pattern extends far beyond the expected network (to nearly all GM).

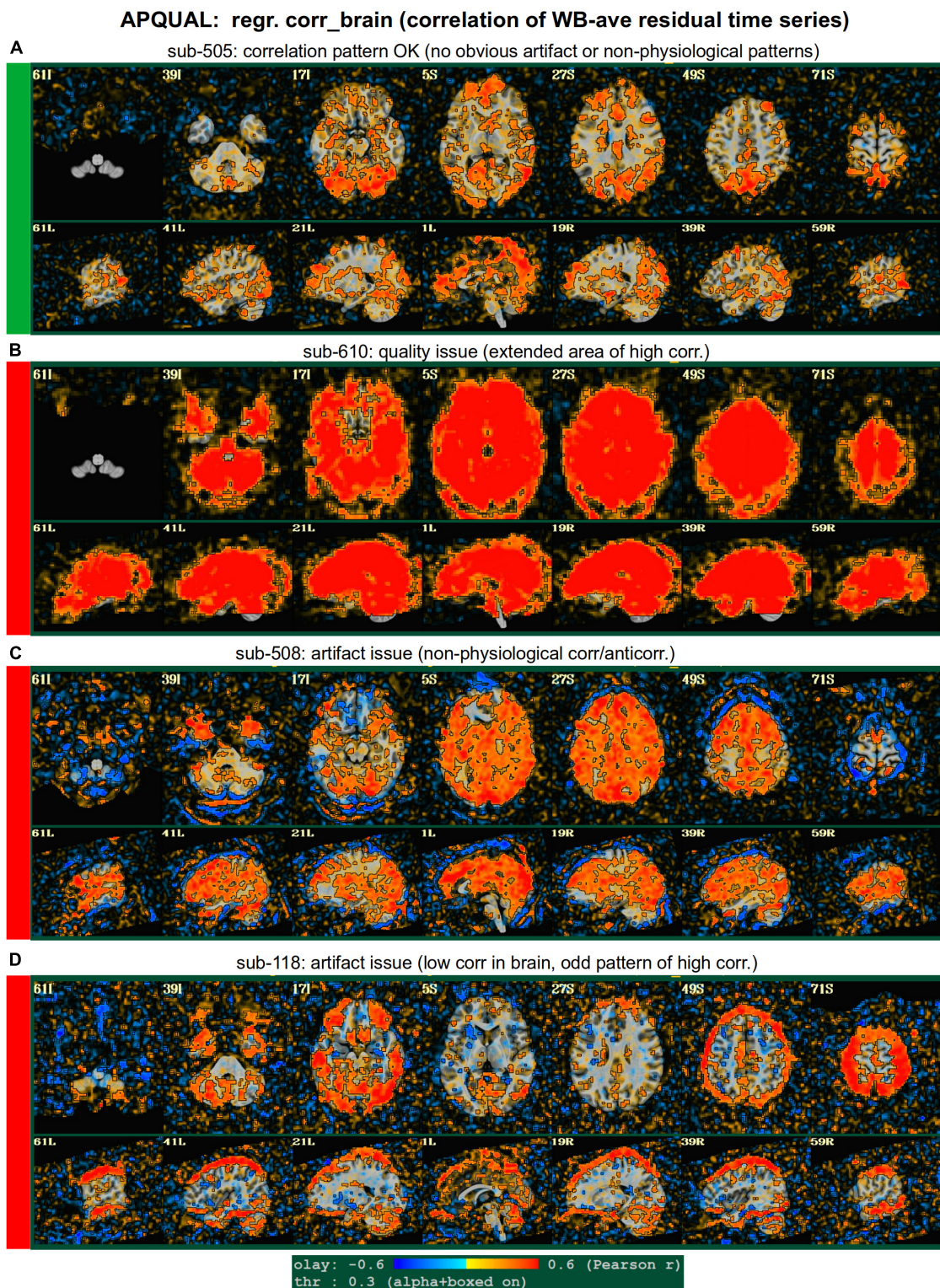


FIGURE 6

APQUAL examples for the “regr” QC block: Regression evaluation through the correlation pattern of the brain-averaged residual time series (“corr_brain” maps). **(A)** Regions of low-medium correlation are mainly located through the GM. **(B)** The whole brain volume correlates highly with the global average, suggestive of strong non-physiological signals remaining in the data. **(C)** High correlation extends through the intracranial regions, with large negative filaments, suggestive of strong non-physiological signals remaining in the data. **(D)** Strong patterns of high correlation remain in the data, outside of GM.

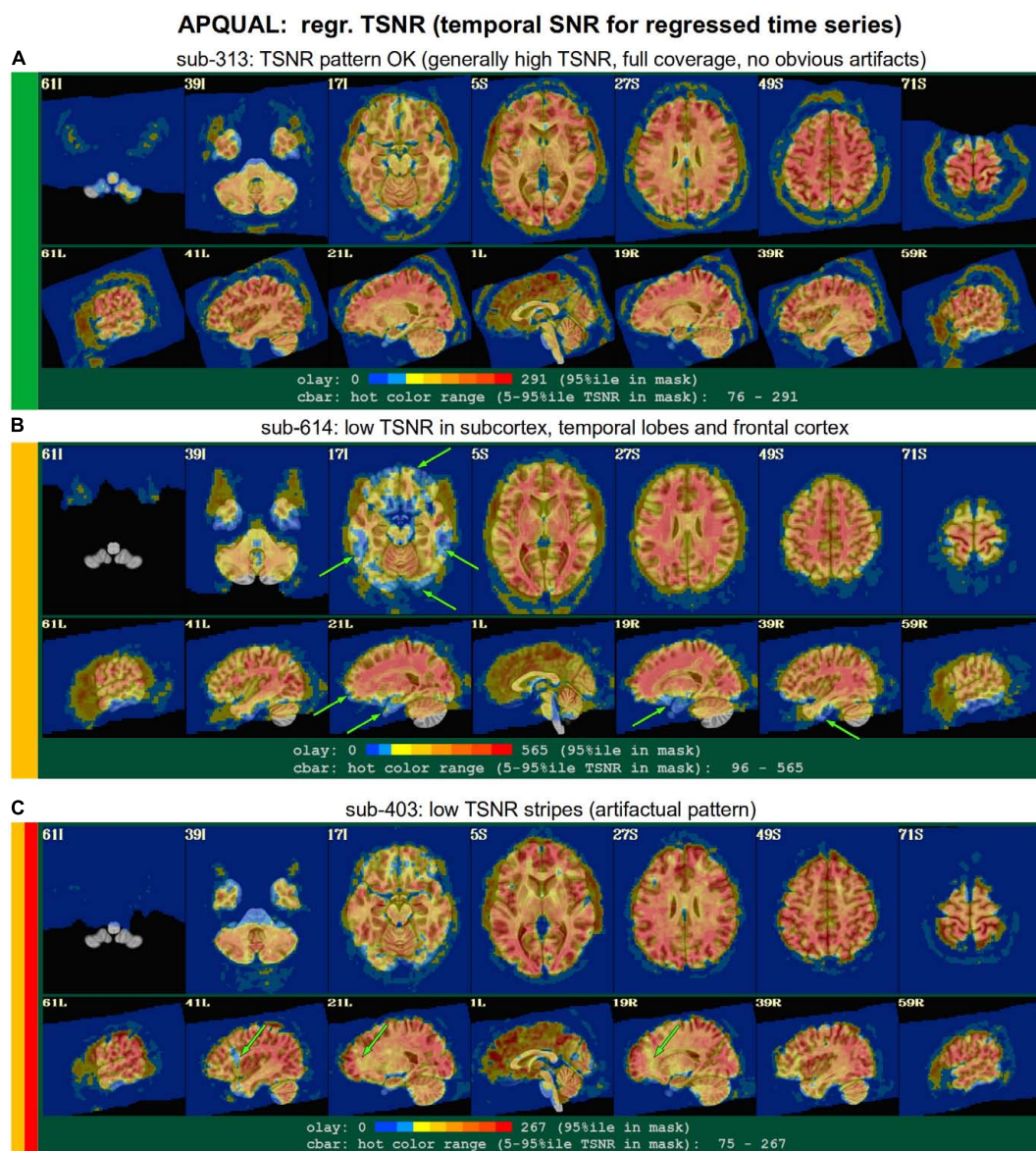


FIGURE 7

APQUAL examples for the “regr” QC block: TSNR maps of the final data after regression modeling (for each voxel, TSNR is the mean of the modeled time series divided by the standard deviation of the residuals). (A) TSNR is relatively constant and high throughout the brain volume (only very small regions of low signal, in the anterior temporal lobes). (B) Large regions of low-TSNR, particularly in the subcortex and orbitofrontal regions, which may impact cortical results. (C) Vertical strips of low TSNR are present, which may affect connectivity analyses (and which, after GUI-based investigation with InstaCorr, appear to be due to a significant artifact, shown in Figure 10, leading to subject exclusion; hence, the inclusion of red in the colorband to the left of the image).

difficult to assess or create an impression of poor alignment. The CSF is particularly bright in Panel D (and for many subjects in Group 7), but the structural alignment still appears to be quite high (albeit in the presence of some geometrical distortions).

vstat.DMN

Figure 4 shows part of the “vstat” QC block, which provides views of statistics based on the regression modeling. For resting state fMRI, this includes seed-based correlation maps when the

final data is in a recognized template space, and the images in this panel use a seed in the left posterior cingulate cortex [L-PCC; coordinate (5L, 49P, 40S) in the MNI template space], which is a standard part of the standard default mode network (DMN) along with medial prefrontal cortex and left/right inferior parietal lobules. This (and the other vstat seed-based vstat maps) provides a useful QC check for noise, artifact and modeling, since generally consistent spatial network patterns appear across age groups, species and alertness/sleep levels.

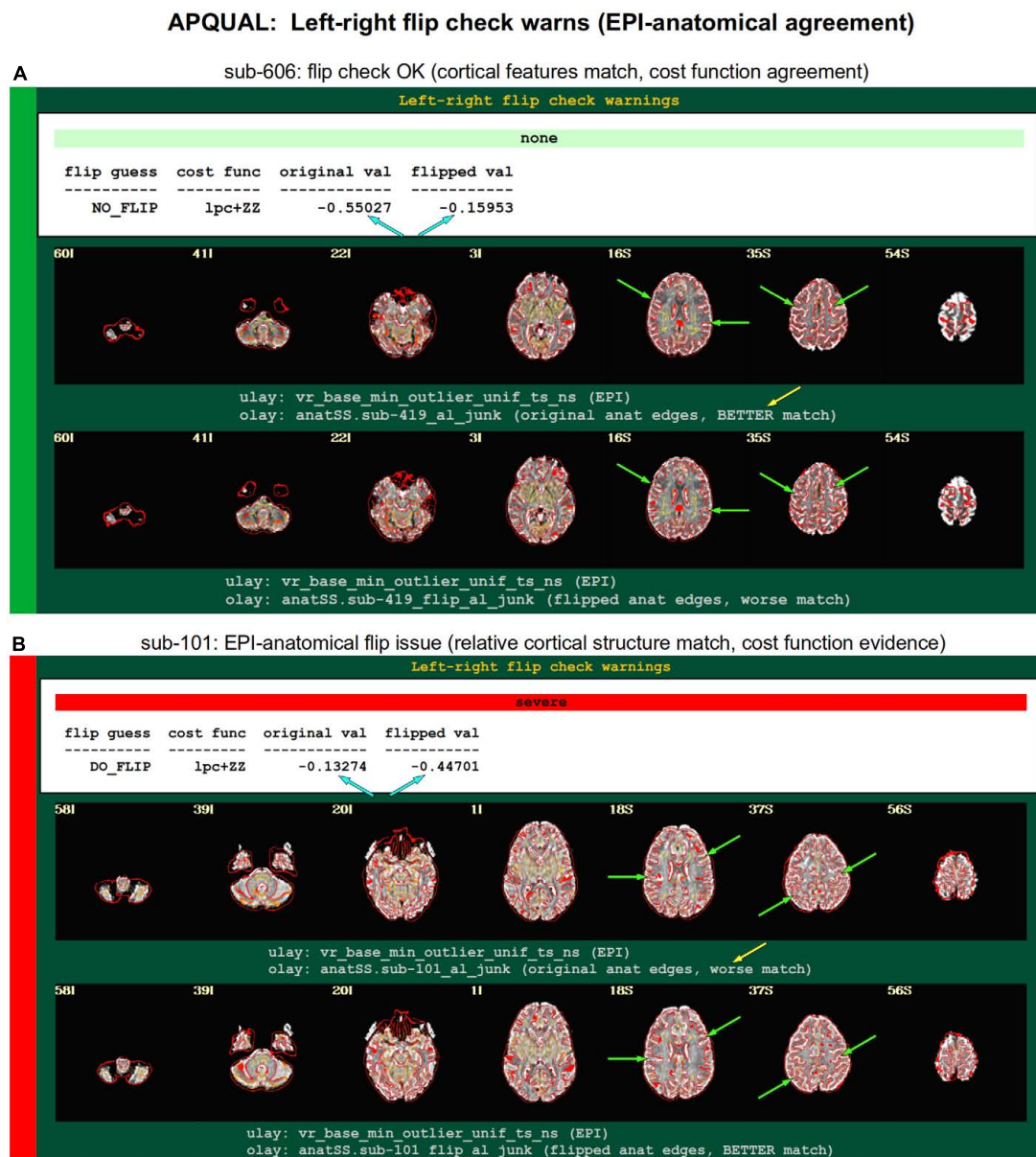


FIGURE 8

APQUAL examples for the “warns” QC block: Warnings created during processing, here for possible left-right flipping between the EPI and anatomical volumes. The warning field contains the APQUANT evaluation, based on cost function comparison (blue arrows), with its comment on the original (yellow arrow) and flipped EPI volumes. Importantly, images of each alignment result within the test are shown, for visual verification of the results. (A) The structures of the original EPI match well with the anatomical volume (and those of the flipped version do not), suggesting consistency. (B) The structures of the original EPI do not match well with the anatomical volume, while those of the flipped version do, suggesting inconsistency in the datasets.

Panel A shows what would be a typically reasonable result for a single subject map for sub-505: The higher correlation regions approximately follow the expected DMN pattern with acceptable specificity and approximate symmetry. Given the generally low SNR of FMRI, as well as length of scanning, one expects small noise patterns of correlation/anticorrelation. Note that here “transparent thresholding” is applied to the overlay, so that results below Pearson $|r| = 0.3$ are still

observed, and brain masking is not applied: These features reduce the sensitivity of results to threshold value and allow for subtle patterns anywhere within the acquired FOV to be observed, which is vital for artifact detection (Taylor et al., 2022).

Figure 4B shows an example of an obvious artifact appearing in the correlation map of sub-203. The slice-wise nature of strong correlation throughout the brain is highly

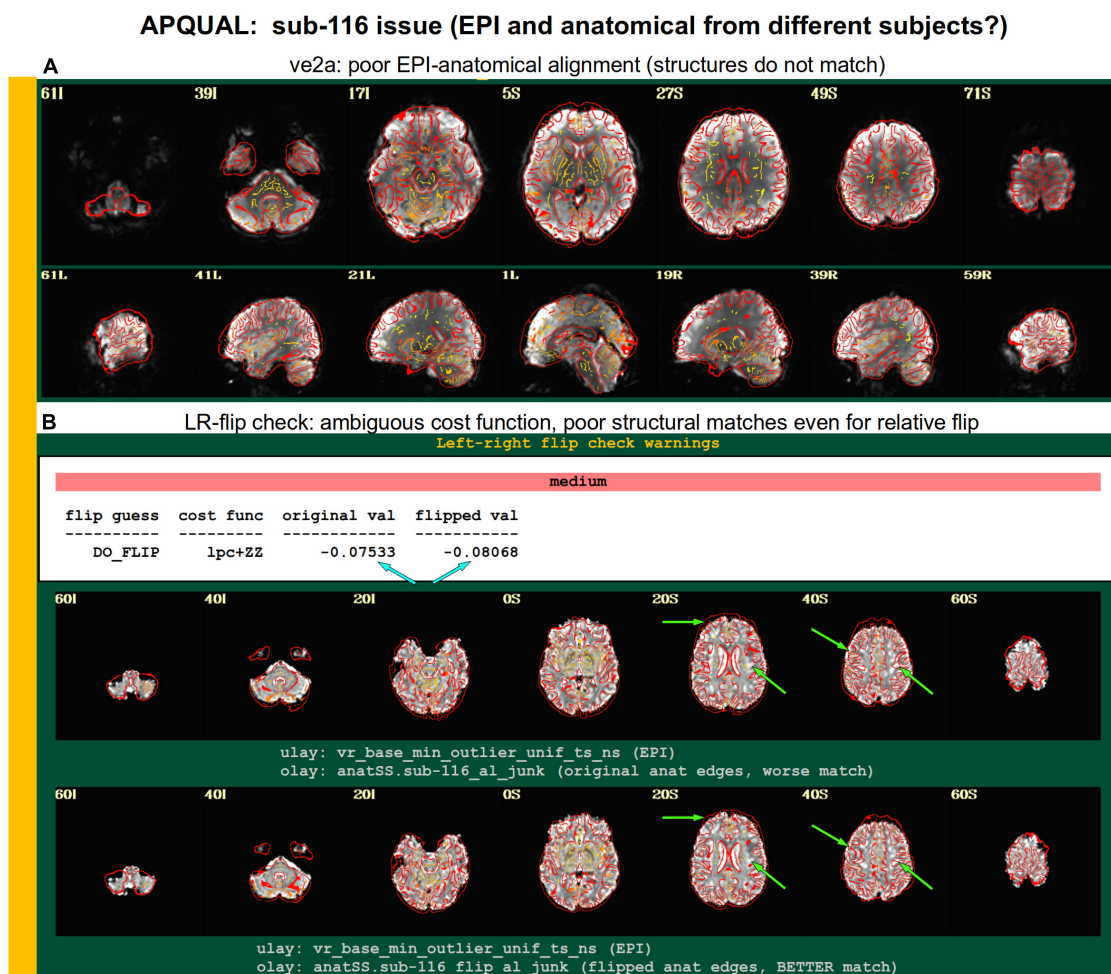


FIGURE 9

Combining APQUAL blocks: ve2a and warns (see Figures 3, 8). The structures of the aligned EPI do not match well with those of the anatomical, even though neither appears heavily distorted [ve2a, (A)]. The left-right flip check provides a “medium” level warning, because the cost function comparison is ambiguous [warns, (B); see blue arrows]. Visually, neither the original nor flipped EPI matches well with the anatomical structures, even though all other subjects in the group had strong alignment. Since the structures appear to differ, this suggests that the EPI and anatomical volumes for this dataset may actually come from different subjects.

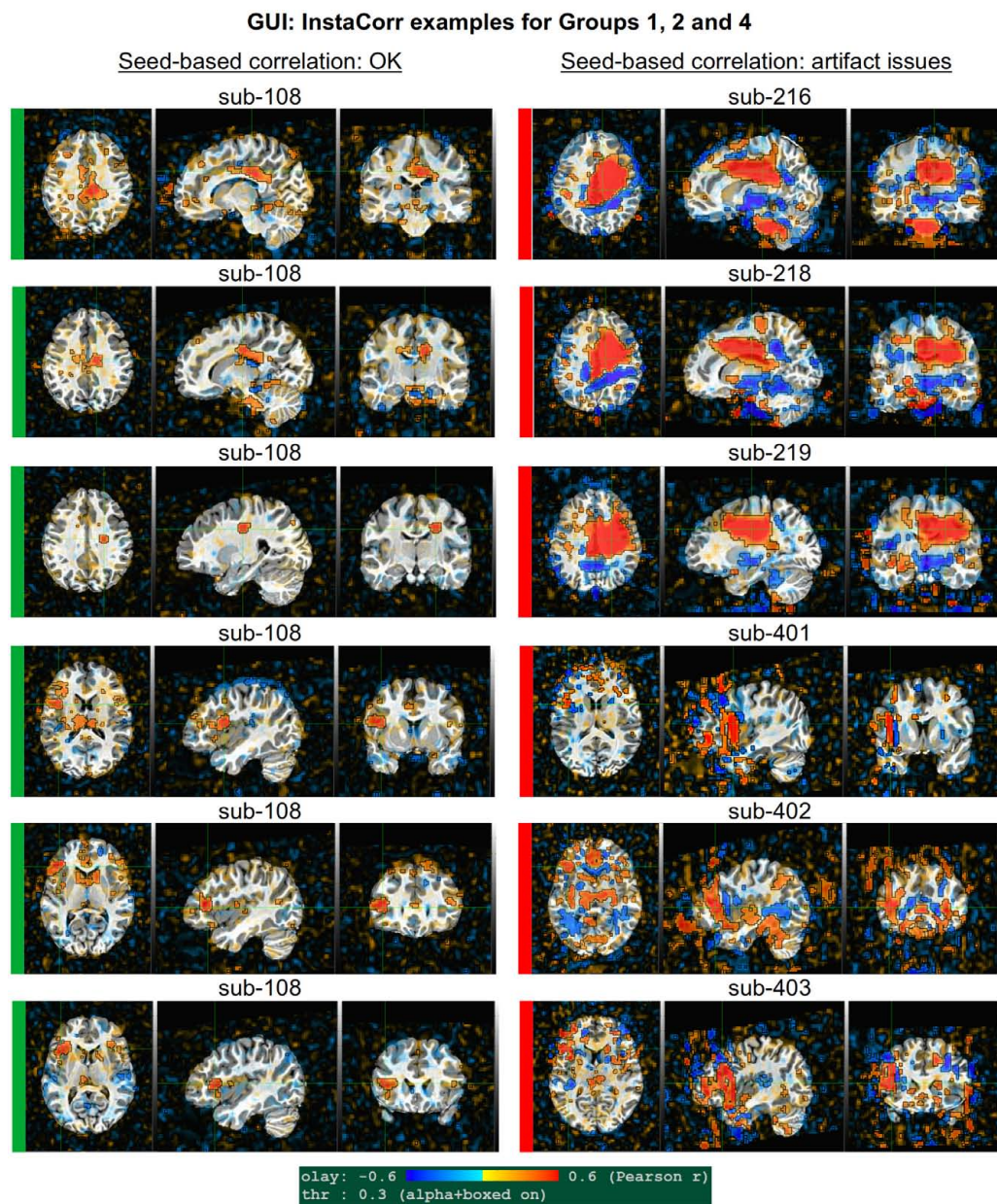
non-physiological, and strongly suggests this subject should be excluded from further analyses. Motion levels and other quantitative QC properties for this subject were not even at a warning level. The other two seed-based maps of the visual and auditory networks did not show obvious artifactual patterns, but the “corr_brain” map and “radcor” maps in the QC did show further extent of odd patterns. As described below, we also applied the GUI to investigate this subject (and others within Group 2), further verifying the presence of artifact (which unfortunately led to the exclusion of all subjects within Group 2).

Panels C and D of Figure 4 show other issues that can be arise in seed-based correlation maps: Noisiness (without an obvious artifact), which includes relatively high correlation/anticorrelation scattered around the FOV and/or

mildly distorted patterns, as for sub-118; and widespread low or missing correlation patterns, as for sub-413 (and alignment quality was verified, so seed location did not appear to be obviously erroneous). In either case, the lack of strong artifact pattern makes it difficult to decide to exclude either subject from these images alone, and further investigations would be needed to avoid biasing the final group selection. (In these cases, the APQUANT stage showed suprathreshold censoring levels of 61% for sub-118, and the GUI-based InstaCorr check revealed notable artifact patterns in the frontal region for sub-413; therefore, from those separate criteria, each subject was excluded).

vstat.vis

Figure 5 shows another vstat visualization, for the visual network [seed located at coordinate (4R, 91A, 3I) in the MNI

**FIGURE 10**

GUI examples of QC, using AFNI's InstaCorr: This provides deeper understanding of the spatiotemporal patterns of the data through interactive driving of seed-based correlation. Several subjects in Groups 2 and 4 had difficult to interpret APQUAL QC results, particularly in seed-based correlation maps (vstat); upon further inspection here, it was apparent that those subjects contained large artifacts within the EPI datasets, as evinced by large correlation/anticorrelation patterns from seed locations in deep WM (Group 2) and extensive, non-physiological correlation/anticorrelation patterns from frontal GM/WM seeds (Group 4). In the end, these artifacts appeared to be present in all subjects of these groups, so that all were categorized for exclusion.

template space]. Panel A shows an expected correlation map for the same sub-505, which essentially contains high correlation in the V1/V2, V3, occipital areas and visual-associated areas. In contrast, panel B shows the presence of large patches of strong correlation and anticorrelation in other parts of the brain for sub-209. Furthermore, these patterns are not constrained by physiological or tissue boundaries. In total, this leads to

excluding this subject (as noted above, GUI follow-up across Group 2 further verified extended artifacts).

In **Figure 5C**, a low/missing correlation pattern is observed for sub-305, again leading to an uncertain evaluation from this image. For this subject, the same low correlation was observed across all seed-based maps, but there was no obvious criterion for exclusion, and therefore the “uncertain” rating remained. In

panel D, sub-403's network map shows unexpectedly extensive regions of high correlation, throughout most of the gray matter (GM). While this "overfull" region of high correlation differs notably from the visual network regions, the lack of distinct, non-physiological patterning makes it difficult to exclude a subject from this image. (This subject's APQUANT criteria were all below threshold, but as noted above, a GUI QC check with InstaCorr revealed that all subjects in Group 4 had a notable artifact within their dataset, leading to their exclusion).

regr.corr_brain

Figure 6 displays another volumetric visualization, which is the "regr" block's "corr_brain" map: The brainwide average of the regression model's residual time series (the "global signal") is correlated with each voxel in the FOV. This essentially provides a visual assessment and corollary to the GCOR parameter (Saad et al., 2013), which is used as a warning and exclusion criterion in the APQUANT QC. Panel A shows a correlation map for sub-505, whose data had generally reasonable correlation maps (and a very subthreshold GCOR = 0.05). Much of the GM shows a generally positive and "medium-level" correlation, with typically low correlation in other tissues. This can be contrasted with sub-610, whose map has universally quite high correlation and leads them to being excluded (as did the associated GCOR = 0.47, in the APQUANT stage).

Figure 6C shows another problematic corr_brain map. While the GCOR = 0.08 for sub-508 is well below threshold, the relatively high correlation patterns across all tissues and anticorrelation boundaries appear to be artifactual. We note that this subject also displayed artifactual patterns in the vstat seed-based correlation maps. The high correlation patterns for sub-118 in panel D do not show the same whole brain coverage, but they do appear to be strongly non-physiological, and lead to this subject also being excluded. (Recall this subject's "uncertain" noisy correlation map in **Figure 3C**, as well as the fact that censoring levels were also at a level for exclusion).

regr.TSNR

In **Figure 7**, TSNR maps for the final, regressed data are shown³. As typical TSNR ranges can vary with scanner site, the colorbar is defined relative to a 5–95% ile interval within the brain mask (providing the min-max values of the hot colors, respectively). Panel A shows a relatively good TSNR pattern: While there is some dropout in the orbitofrontal regions and temporal lobes for sub-313, such effects are present in nearly all fMRI and the TSNR strength is relatively constant across the brain and GM. If the low TSNR is not in a focal region of the study, then this subject would be fine to include in the subsequent analyses; for studies that include these regions

of typical signal loss, one would have to adjust acquisition parameters to avoid problematic distortions. (Note that one can observe the tight FOV for this subject's EPI, which would preferably be larger to avoid TSNR issues in the superior slices, as well).

The TSNR map for sub-614 in **Figure 7B** shows a larger area of dropout in the inferior regions of the brain. As shown in the images, a larger fraction of the temporal lobe, subcortex and orbitofrontal regions have notably lower TSNR than the rest of the brain. As whole brain connectivity studies often include these regions, it is likely that such differences in fMRI signal could affect the final results, depending on the hypotheses and exact paradigm. Therefore, this subject may not be appropriate to include in the study, and is rated "uncertain" from these images.

Figure 7C shows a TSNR map for sub-403 with relatively full whole brain coverage of constant TSNR, even in the inferior and subcortical regions. However, there are notable vertical stripes of low-TSNR that appear in each hemisphere in the anterior regions (see the sagittal slices). Such non-physiological patterns suggest some kind of artifactual signal issue, such as significantly strong ghosting, which may affect large areas of interest. Therefore, these patterns may mean that this subject would be inappropriate to include in further analyses. However, we note that in a follow-up QC analysis using InstaCorr in the AFNI GUI, these striped locations showed extreme and non-physiological patterns of correlation/anticorrelation (described further below, and see **Figure 10**). These low TSNR stripes were observed across Group 4, and the GUI follow-up revealed the same artifact in all subjects, leading to the exclusion of this group. Thus, in this group the low-TSNR striping was a hallmark of an artifact that always led to excluding a subject, but it is possible that in other datasets, that might not be the case. At the least, such patterns warrant detailed follow-up, likely using the GUI.

warns.flip

The APQC HTML contains a "warns" section that is comprised of the results of various automatic checks that occur during afni_proc.py processing (see list, above). Each has an associated warning level of "none," "mild," "medium," "severe," or "uncertain." **Figure 8** shows the results of a particular warning that spans the APQUANT and APQUAL QC: Checking for left-right flips between the EPI and anatomical volumes. Panel A's results suggest that sub-606 does not show an inconsistency: The cost function value of the original data set is much lower than the flipped version (blue arrows; and note that cost functions are minimized in the alignment process), and the images below allow one to visually verify that the cortical patterns of the original EPI are much more consistent with those of the anatomical volume. NB: The structures of the superior cortex

³ TSNR can be variously defined in fMRI studies. Here, TSNR is the ratio of the mean of the voxel's final time series to the standard deviation of its residual time series.

tend to be much less left-right symmetric than the inferior regions and subcortex, and therefore provide more convincing evidence.

Figure 8B shows an example of the quantitative flip-check strongly suggesting that sub-101's original EPI and anatomical volumes have a relative left-right flip. This result is visually verifiable in the associated images. Since the *absolute* left-right definition cannot be known (without external indication such as a vitamin E tablet in the FOV), this subject would be excluded from further analysis. The data for sub-115 in this group similarly appeared to have a left-right flip.

A particularly interesting case of left-right flip check results is shown in **Figure 9**. Here, sub-116's ve2a check initially showed a relatively poor EPI-to-anatomical alignment. Additionally, the left-right flip check provides a "medium" level warning, because the cost function values when using the original or flipped EPI are extremely close; in such a case, the recommendation whether to flip or not is difficult to interpret, as it is effectively "within the error bars" of the alignment's cost function estimation. Looking at all of the images, it appears as if the cortical structures of the EPI and anatomical volumes do not match well in either case. Given that the EPI distortion is not very large and that the EPI-anatomical alignment for all other subjects from the site displayed excellent structural correspondence, these QC results suggest that the two volumes in sub-116's dataset did not actually come from the same subject. When using a publicly downloaded dataset, this is only a supposition and cannot be directly verified, and, therefore, we are uncertain about whether to include this "subject" in further analyses.

GUI evaluation with InstaCorr

The APQUAL and APQUANT items listed above provide useful QC information: The quantitative and visual aspects provide complementary aspects for efficiently and systematically understanding many aspects of the data. For example, the EPI and anatomical left-right flip check can be quantitatively evaluated, but should always be visually verified. As shown for sub-116 in **Figure 9**, data visualizations are sometimes even necessary for interpreting quantitative findings appropriately. However, in some cases even the APQUAL visualizations did not contain enough information to confidently make a QC evaluation. Therefore, the GUI stage of QC was used in several cases, in particular using the "run" script provided by `afni_proc.py` to efficiently start the AFNI GUI with InstaCorr set up, to explore the spatiotemporal properties of the EPI data.

Figure 10 shows a set of representative GUI snapshots when applying InstaCorr. As noted above, some of the correlation patterns for subjects in Groups 2 and 4 were not as expected: Some contained large patches of correlation and

anticorrelation; some contained faint (subthreshold) patterns that were difficult to interpret; some contained extremely low or missing spatial patterns. For all subjects in these groups, the GUI follow-up revealed strong artifactual patterns in seed-based correlations, and example of these are shown for a subset of each group and contrasted with what might be considered a reasonable pattern at the same location in subject that did not appear to have artifacts (sub-108).

The seed location for each of the Group 2 subjects (sub-216, sub-218 and sub-219) is located in deep white matter (WM), which should have minimal patterns of correlation. As in the left column, one might expect a small, local patch of correlation even in WM, due to data blurring, remaining motion artifacts, vascular-driven BOLD response in WM, and more. However, the large patterns of high correlation/anticorrelation for each Group 2 subject spans tissue boundaries non-physiologically. Since these patterns overlap variously with GM, they do not appear possible to separate typical resting state connectivity analyses, and therefore all of Group 2's subjects were categorized for exclusion.

InstaCorr analysis for Group 4 (sub-401, sub-402, and sub-403) revealed a different location of artifact, as shown in the lower panels of **Figure 10**. With a seed located in either the left or right frontal GM or WM, again strong patterns of high correlation and anticorrelation appeared, in this case alternating and even extending outside the brain. Again, these patterns are in contrast with expected local and/or localized symmetric patterns of correlation, depending on the GM/WM content of the seed region. These artifactual patterns throughout the frontal cortex GM also imply that resting state connectivity analyses would be strongly affected by non-physiological features, and therefore all of Group 4's subjects were also categorized for exclusion.

Group summary QC notes

After performing the detailed single subject checks listed above, it can be useful to summarize features or trends that appear across the group. These may be helpful for judging the overall applicability of a data collection for a particular study question. Additionally, these may aid planning future studies, by either replicating important features or by avoiding non-ideal aspects, possibly adjusting acquisition parameters. In general, the following overall properties of each group are based on the visualization methods described in the APQUAL stage.

Group 1 had several subjects with relatively low visual cortex correlation in `vstat.vis` seed-based correlation maps, even though the other network correlation maps were more standardly represented. There were some light vertical striping

patterns in the frontal brain regions of the TSNR plots, suggesting some mild ghosting effects. Finally, in the individual motion parameter plots, the dP (translation along A-P axis) tended to have a noticeably linear increase across time, which might be due to frequency drift (e.g., Foerster et al., 2005) or even from settling into a pillow; while not necessarily a problem, this is an example of a group-wide feature in the data that is worth understanding, particularly if acquiring one's own data.

Subjects in Group 2 had relatively high corr_brain maps, and the TSNR dipped noticeably in the center of the brain. The radial correlation (radcor) patterns were noticeably high centrally, and this led to discovering the presence of a strong artifact across all subjects, using InstaCorr. If subject data were still being acquired, such an artifact might encourage close examination of all datasets coming from that particular scanner.

Groups 3 and 4 each had relatively tight FOV for the EPI acquisitions. These might negatively affect signal quality in some boundary regions.

Group 4's EPI volumes had quite short time series (123 points). The TSNR plots showed a strong vertical striping pattern, which led to the discovery of a notable frontal artifact across all subjects, using InstaCorr. The motion plots revealed a steady dP translation over time (as well as some notable linear trends in other parameters).

In Group 5, the basic acquisition features of voxel dimension and matrix size were quite heterogeneous. Non-linear alignment of the highly anisotropic EPI voxels (1.87 mm × 1.87 mm × 4.0 mm) produced slight swirls in patterns, which is one reason that acquiring anisotropic voxels is not recommended for standard group analyses; it also creates a grid-based dependence for the acquired data (e.g., which brain regions are averaged together depends on the orientation of a subject's head in the scanner), a property that should be avoided. There was also noticeable signal loss in the orbitofrontal and temporal lobes, as well as the subcortex, which may lead to the exclusion of most of these subjects in some whole brain studies, depending on the specific regions of interest.

Group 6 also had a large heterogeneity in basic acquisition parameters, particularly in terms of number of EPI runs and run lengths, as well as matrix sizes. There was notable geometric distortion in the EPIs, particularly along the phase encode axis, with both signal pileup and attenuation; due to the different patterns of distortion, the phase encode direction may have been inconsistent across the group. TSNR was high across much of the brain, but low in the orbitofrontal and temporal lobes. There were relatively high values of the corr_brain (the correlation of the average residual signal across the brain).

Group 7 had notably bright CSF in the frontal portions of the brain in the EPI, but this did not appear detrimental to alignment or analyses. This group seemed relatively prone to motion, with many subjects having unusually high censor fractions.

TABLE 4 Summary of the first stage of task-based FMRI QC: GTKYD ("getting to know your data").

GTKYD: "Getting To Know Your Data" results (task-based FMRI)	
Property	Description
Group 0: EPI	
orient diff	sub-010 has RIA, from group std RPI
oblique	
anatomical	
(some) oblique	

This displays cases of heterogeneity in basic dataset properties, as well as noteworthy values for checking or for informing processing choices. Items shown here might prompt verification with the source of the data collection, whether it has been downloaded from a shared repository or is being acquired locally.

Results for task-based data collection

GTKYD summary

Similar to the analysis of resting state FMRI, GTKYD was the first stage of checking each group's data, and no subject exclusions were made from this step. The summary of basic dataset properties for the single group of task-based FMRI (Group 0, 30 subjects) is shown in Table 4. One subject's EPI had a different orientation from the rest of the subjects. While all EPI volumes were acquired obliquely, only a subset of anatomical volumes were acquired obliquely.

The table of GTKYD checks for the task-based FMRI group is shown in Table 4. Here, one subject's EPI had a different orientation than the rest. While all EPI volumes were acquired obliquely, only some of the anatomical volumes had obliquity information; as with the resting state data, we chose to deoblique these anatomicals as an initial processing step. Finally, no slice timing information was present for these EPI volumes.

APQUANT

Table 5 shows a brief summary applying APQUANT exclusion criteria (itemized in Table 1) and additional APQUAL and GUI checks to the task-based FMRI group. The same subject dataset categorizations (described above): Include, Exclude and Uncertain. The Supplementary Table 1 contains a table with more detailed descriptions for each subject.

The task-based FMRI data from 30 total subjects were processed. Following the QC checks, 15 were categorized to include for further analysis, 7 to exclude and 8 were listed as uncertain. Each excluded subject had at least one APQUANT criterion that resulted in that categorization (and typically multiple ones, as well as APQUAL items; see the detailed Supplementary Table 1). Most of the "uncertain"

TABLE 5 A brief summary of task-based FMRI dataset evaluations, based on the APQUANT, APQUAL and GUI QC checks.

QC evaluations (brief): Group 0 (task-based FMRI)				
Group 0 (I = 15, E = 7, U = 8)				
sub	eval	comment		
001	I		016	U APQUAL.vstat.quality
002	I		017	E APQUANT.excl('fraction TRs censored')
003	I		018	I
004	I		019	I
005	U	APQUAL.vstat.quality	020	U APQUAL.vstat.quality
006	I		021	U APQUAL.vorig.EPI
007	I		022	E APQUANT.excl('fraction TRs censored')
008	I		023	U APQUAL.vstat.quality
009	E	APQUANT.excl('fraction TRs censored')	024	E APQUANT.excl('fraction TRs censored')
010	U	APQUAL.vstat.quality	025	U APQUAL.vstat.quality
011	I		026	E APQUANT.excl('fraction TRs censored')
012	E	APQUANT.excl('fraction TRs censored')	027	E APQUANT.excl('fraction TRs censored')
013	U	APQUAL.vstat.quality	028	I
014	I		029	I
015	I		030	I

The following abbreviations for evaluations ("eval") are used: E, exclude; I, include; U, uncertain. A more detailed summary is provided in the [Supplementary Table 1](#), with further comments about most subjects.

categorizations were due APQUAL examination, particularly to visualization of the statistical results, which are described in the next section.

APQUAL evaluation

Figures 11–14 contain example images of the APQUAL evaluations for Group 0. These figures come from the APQC HTML report, of which most QC blocks are the same as for resting state FMRI. One exception is the vstat block, which shows F-stats and modeling coefficients (effect estimates) and associated statistics. The same colorband labels used for the resting state examples (see [Figure 2](#)) are used, as well as arrows to highlight features of note. In general, there were fewer QC issues with this group than for Groups 1–7. Therefore, we focus on different features in the overlapping blocks, as well as some of the stimulus-specific QC considerations.

vorig

Figure 11 shows QC examples from the “minimum outlier” EPI, used as a reference for motion correction and anatomical alignment. In panel A, sub-030's volume does not display any obvious artifact or major distortion. The tissue contrast is also reasonable (some of the superior slices have slightly higher brightness, but the maximum value did not show saturation). In panel B, the FOV is much tighter for sub-020, and there is a notable ghosting artifact: The brain and skull from the

posterior part of the brain is wrapped around to the anterior, and here appears to overlap with the brain volume. While different degrees of ghosting occur in many EPI acquisitions, there is a question here of whether the visible overlap suggests problematically strong signal interference in a non-negligible region of the study. In panel C, the inferior slices show the presence of ghosting or phase artifact. The distortion is limited to approximately the bottom ten slices, but this includes large portions of the frontal and temporal lobes (as well as other parts of the brain).

vstat

Figure 12 shows images of the full F-stat maps, which are part of each “vstat” block for task-based FMRI. As a ratio of explained variance to unexplained variance after regression, the full F-stat provides information on the relative model fit (higher values = better fit). These images provide a useful QC check for noise, stimulus modeling and motion reduction, though their details and expected patterns will (necessarily) vary strongly by paradigm. While there might be some expectation of regions of high F-stat that should be observed (e.g., the visual cortex when an on/off stimulus is presented visually; the motor cortex when button responses are used; some particular region from a previous study or theoretical rationale), it is difficult to apply an unexpected patterns as a drop criterion, unless an obvious artifact is observed, for example. Strong deviations across many subjects may be a sign of study design issues, subject unresponsiveness, stimulus timing issues or simply unexpected

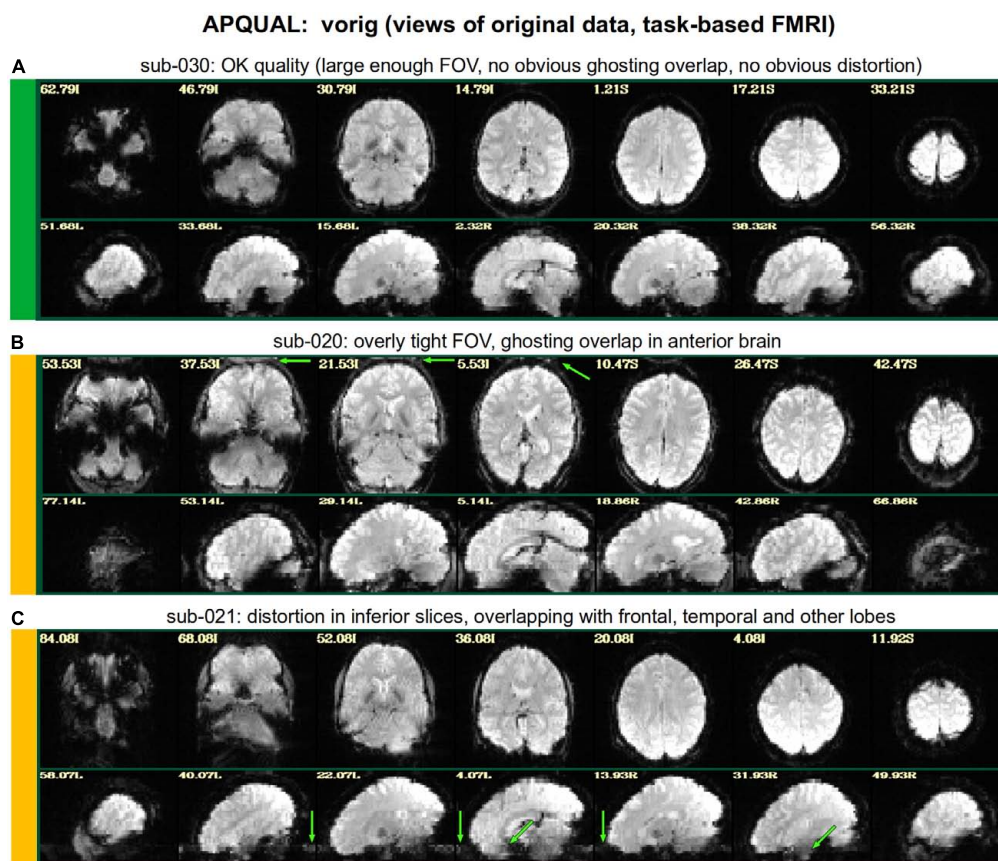


FIGURE 11

APQUAL examples for the task-based FMRI group from the “vorig” QC block: Visualizations of the original datasets (here, just the EPIs). See [Table 5](#) for brief, overall evaluations for each subject, and the [Supplementary Table 1](#) for detailed QC comments. (A) The EPI does not appear to have any major artifact, ghosting or distortion, and tissue contrast is reasonable. (B) The FOV of this volume is overly tight for this subject, so that there is ghosting of the posterior brain and skull which overlaps the anterior portion. (C) The inferior slices show a ghosting or phase distortion artifact—part of the frontal and temporal lobe regions are notably distorted.

findings. While worth noting and commenting on, variations in statistical patterns will still be expected, and one must be careful not to bias results in the QC process.

In the present study, panel A of this figure (sub-001) shows what is likely a reasonable quality F-stat map for the present paradigm. A similar F-stat range (99% ile within the brain mask >40) and spatial pattern [high values in visual cortex, and left and right inferior frontal junction (IFJ); see green arrows in $Z = 27S$] were observed across many subjects, particularly among those with no obvious exclusion criterion. The high F-stat regions are localized in GM, and no obvious artifact or non-physiological patterns are observed.

Panels B and C show two subjects (sub-005 and sub-016, respectively) with generally lower F-stat values across the brain (99%ile within the brain mask <10). Note that motion and censoring levels for these subjects were not particularly, and no quantitative (APQUANT) criteria suggested excluding them. In the vstat images, the relative noise levels are higher and observed throughout the intracranial region, and there are fewer

obvious patterns of localized clusters of high F-stat. In B, relatively high F-stat clusters appear in the IFJ, but are barely observable in the visual cortex; in C, the opposite is the case, with the ventricles also showing surprisingly high F-stat. Such variations from the “standard” pattern are difficult to interpret, but are worrisome for including these subjects in group analysis. Further exploration was made using InstaCorr in the GUI (described below).

One of the additional vstat images automatically created was for the “TASK” stimulus, which is shown in [Figure 13](#). This shows the coefficient (effect estimate) for the stimulus as the overlay, which here has units of BOLD% signal change, scaled to a 2 s stimulus, due to the inclusion of the scaling block and typical mean stimulus durations. Observing the coefficient (instead of just overlaying the statistic itself) is useful for interpreting the model results and judging their reasonableness. The stimulus (and contrast) plots also contain useful sign information, which is lacking in the F-stat images. Here, the locations of large effect and statistical significance

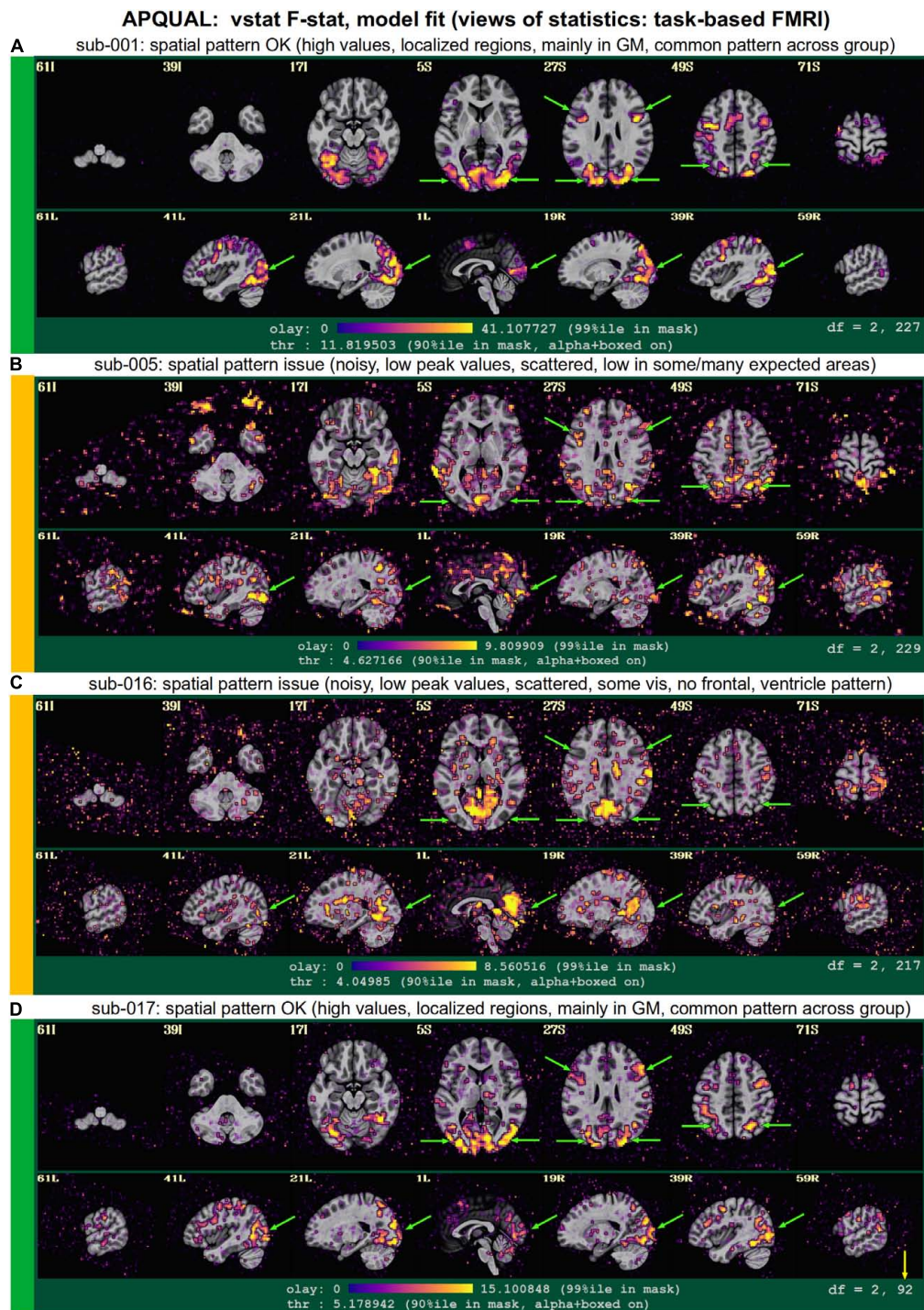


FIGURE 12

APQUAL examples for the task-based FMRI group from "vstat" QC block: Visualizations of statistical and modeling information after regression (here, the full F-stat from the regression modeling, highlighting regions of high model fitting). (A) High F-stat values are localized in GM (esp. visual cortex, and perhaps some in expected regions, if background knowledge is present), and this spatial pattern is fairly typical across the group (green arrows). (B) Compared to (A) the F-stat values are much lower (poorer fits) and less localized in GM, including the visual cortex, though the frontal regions in slice Z=27S are observable; scattered noise has relatively high amplitude. (C) Compared to (A) F-stat values are much lower (poor fits) and less localized in GM, though part of the visual cortex is observed clearly; the ventricles have relatively high F-stat. (D) This dataset has similarly reasonable properties as dataset A, even though 57% of its time points were censored due to motion (note the second value in the degree of freedom count, $df = 92$, is much lower than the other volumes); this subject was still excluded, because of the automatic quantitative (APQUANT) criteria.

typically mirror the high F-stat locations for panels A–D. Note that in panels B, the IFJ regions do not appear to have very strong “Task” stimulus response (relatively low magnitudes and statistics values). In panel C, the ventricles (which had high F-stat values in the same panel of [Figure 13](#)) show negative coefficients for this stimulus. While these images provide further useful details, again we note that the GUI was used to provide further information for sub-005 and sub-016.

vstat, mot, regr

[Figure 14](#) shows several QC block results for sub-024. The vstat image in panel A shows a noisy statistical pattern and overall low peak F-stat values. Looking at other QC blocks or data aspects may provide useful information about why this dataset looks different, such as: Subject motion, lack of stimulus response, mismatched timing files, acquisition artifact or something else. This insight may be particularly important if checking datasets as they are acquired, to determine if study design or setup may be leading to a higher chance of having poor quality datasets.

For this figure’s sub-024, 36% of the time points were censored during processing (as well as >34% of each stimulus class’s response time), and the Enorm and outlier fraction plots (with threshold values and censoring bands) are shown in panel B. This high censor fraction led to this being categorized to be excluded in the APQUANT section, both because of the large information loss during stimulus events and due to the likely presence of remaining motion effects in the non-censored time points in practice; however, some subjects with high censor fractions do have stimulus response maps that appear to have reasonable quality (see panel D of [Figures 12, 13](#)), particularly if the motion is not strongly linked to stimulus events. Panels C and D show the ideal BOLD response curves for this subject, for both the individual stimuli and their sum, respectively, which also contain the censoring bands for reference. In this case, one might observe a possible trend of censoring during or immediately following stimulus events: It is possible that this subject has stimulus correlated motion, so that regression out motion regressors would also remove much of the stimulus-specific features. If several subjects contained such a correlation, then this would suggest the study design should be adjusted, or further procedures taken to reduce motion (e.g., giving specific instructions for the subjects, or having subjects practice the task and then provide feedback if motion appears high). Further QC investigations using an interactive GUI are described in the next section.

GUI evaluation: InstaCorr

Following the APQUANT and APQUAL stages described above, we further explored several of the datasets using the GUI,

again using the “run” InstaCorr script provided by `afni_proc.py`. This can be useful generally to observe artifacts or systemic spatiotemporal features in the data. In particular, the APQUAL reports showed most subjects having strong task responses in visual areas, while others did not, some even when motion was low. This prompted a review using InstaCorr, which showed multiple features. [Figure 15](#) shows InstaCorr images from sub-001 and sub-005 as respective examples of having strong task responses and not. While sub-005 had a poor task response, there were high correlations in the visual area (top row) and IJF (second row), akin to those of sub-001. Were we collecting this data locally, we would review the stimulus timing file creation, to be sure there were no mistakes. But sub-005 also shows unusual correlation and anti-correlation patterns between GM and deep WM, as well as with the ventricles. This led to the “uncertain” QC evaluation of sub-005.

STIM evaluation

All subjects had essentially the same event onset timing, within 0.1 s, except for 2 subjects (sub-002, sub-026) for whom all events started 2 s later. Onsets (ignoring stimulus duration) were separated by times from 2.5 up to 18.5 s, with a mean of 7.5 s and a standard deviation of 3.5. When response time was applied for the duration, Control events had per-subject duration means from 0.51 up to 1.57 s, with an overall range of ≈ 0.0 –2.43 s. Task events had per-subject means from 0.45 up to 2.65 s, with an overall range of ≈ 0.0 –4 s (with the latter being the maximum possible). ISI times (onset separations minus stimulus durations) ranged from 1.3 to 17.3 s, with a mean of 6.4 s. With well separated events, there were no concerning pairwise correlations between regressors. Though we note that since there were only two conditions, they were mildly predictive of each other, leading to typical pairwise correlations around -0.4 for those regressors of interest. The regression matrix condition numbers (computed as the ratio of the largest to smallest eigenvalues) very modestly ranged from 41.6 to 325.0, and were that high only due to correlations among the motion regressors.

Group summary notes

The EPI volumes for Group 0 tended to have a tight FOV, particularly along the anterior-posterior brain axis. For several subjects, the strength of ghosting was large enough to be observed overlapping the frontal brain regions, which can create artifacts. There was notable EPI distortion in the inferior slices of several subjects, and the TSNR was generally low in the subcortex, temporal lobe and orbitofrontal lobe. Additionally, nearly every subject had the same timing onset; it is more common in newer studies that subjects would have randomized stimulus timing, though with similar statistical properties.

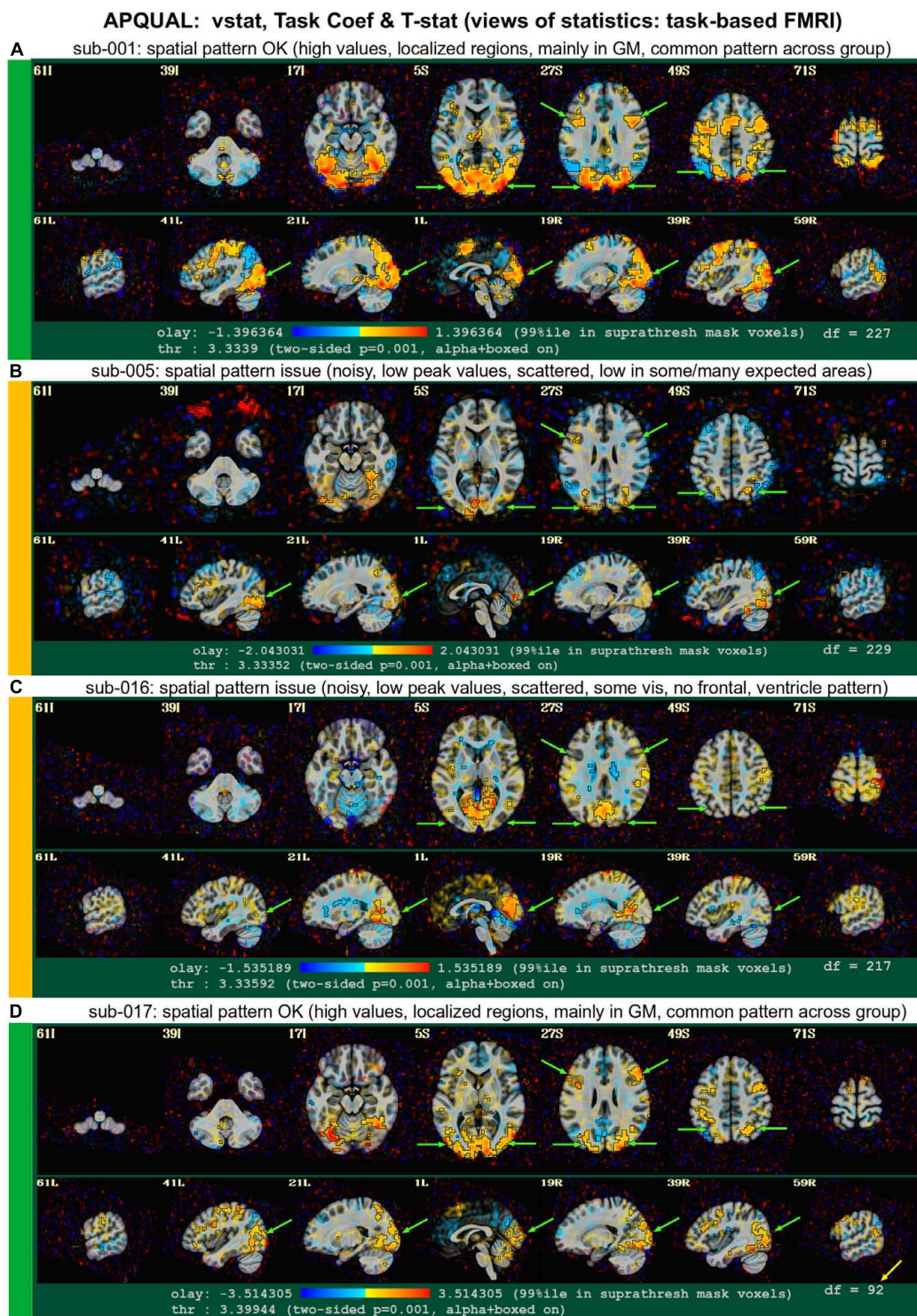


FIGURE 13

APQUAL examples for the task-based FMRI group from "vstat" QC block: Visualizations of statistical and modeling information after regression (here, the "TASK" stimulus coefficient is shown as the overlay colors, and its t-statistic values are used for thresholding). Each panel corresponds to that of Figure 12, though a different aspect of the modeling is shown here: Namely, the task stimulus coefficient that, after scaling, now has physical units of BOLD percent change, as well as the associated statistic (used for thresholding). Similar comments generally apply for each subject to those of Figure 12, but note that: For sub-005 [panel (B)], the high F-stat regions in frontal regions in $Z = 27S$ were not strongly associated with this task, unlike in panels (A,D); and for sub-016 [panel (C)], the ventricle pattern noted in the previous figure are negatively associated with the main task.

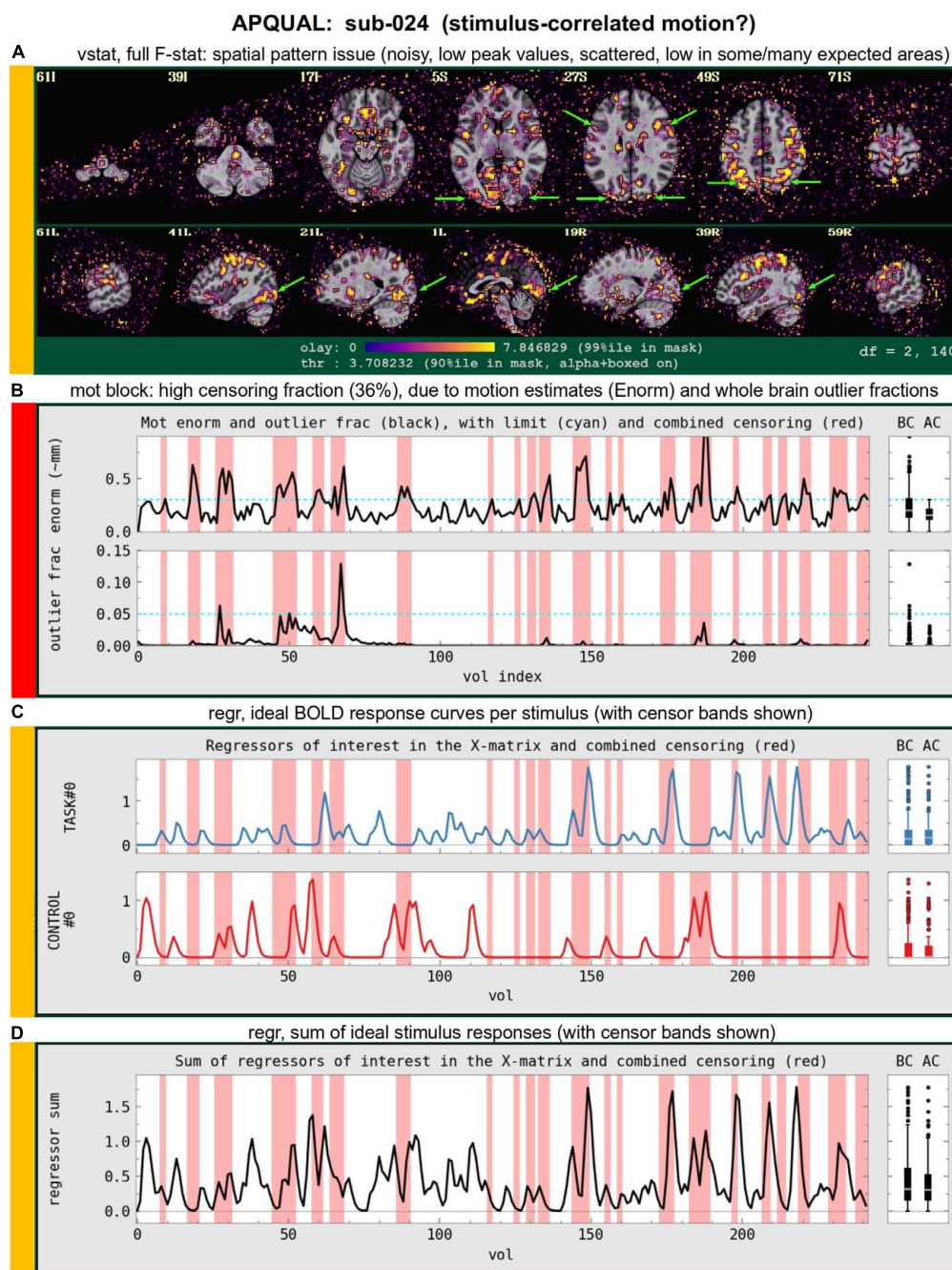


FIGURE 14

Combining APQUAL blocks: vstat (see Figure 12), mot (for combined motion estimates and censoring), and regr block plots of stimulus responses. The full F-stat map for sub-024 in panel (A) is noisy and shows relatively poor model fitting across the brain (cf Figures 12A, D). In trying to understand more about this subject's data, the motion estimate responses are shown in panel (B), where a large fraction of time points have been censored (> 36%; shown in the red bands). Furthermore, in viewing the locations of censoring with respect to the ideal stimulus response curves for this subject [panels (C,D)], one sees that much of the motion appears to occur during many of the stimulus events. Thus, it is possible that this subject exhibits stimulus-correlated motion, which is particularly difficult to remove with modeling.

Discussion

We have described a multi-stage process of QC for FMRI datasets. The stages are layered and complementary to help

researchers understand their neuroimaging datasets, which themselves are complex and require many levels of processing that should be verified. We also introduced a standardized ontology to organize the recording and reporting of the QC

GUI: InstaCorr examples for Group 0

Investigating subjects with quality of vstat maps: surprisingly similar seed-based corr maps?

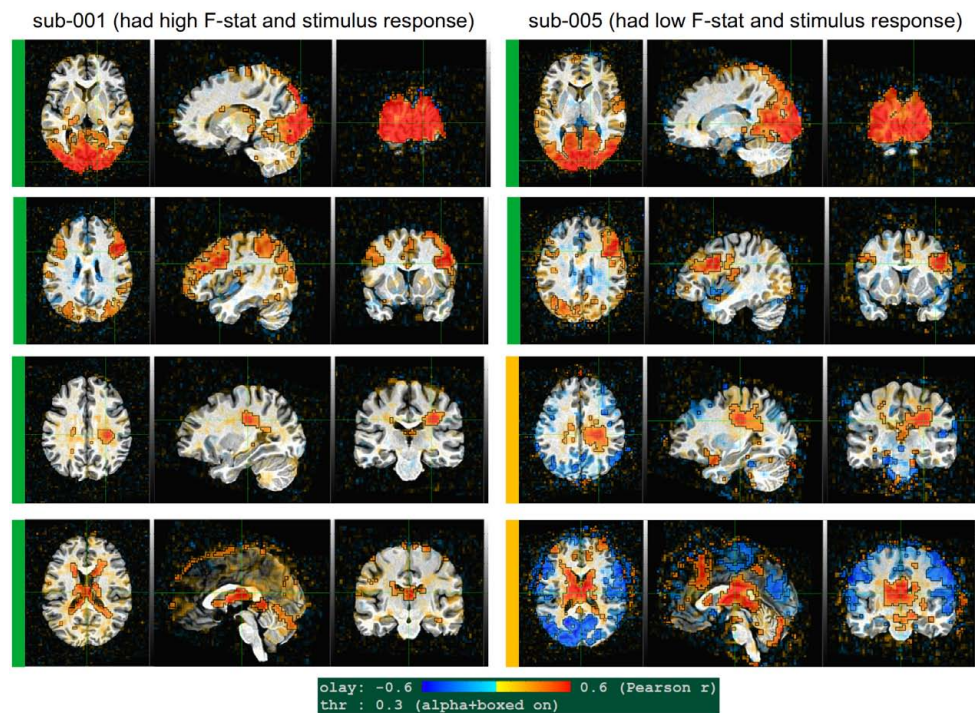


FIGURE 15

GUI examples of QC for task-based fMRI, using AFNI's InstaCorr, to explore the spatiotemporal patterns of the EPI residuals with interactive seed-based correlation. The crosshairs show the seed location for two different subjects: sub-001 (left col) and sub-005 (right col); the same seed location is used, per row of images. The positive correlation responses are quite similar in both supra-threshold spatial coverage and magnitude for seeds in the visual cortex and IFJ (top and second rows, respectively). However, large, unexpected anticorrelation patterns in GM were observed for sub-005, leading to this subject being evaluated as "uncertain."

procedure. These QC methods have grown and adapted over time, and will surely continue to do so, particularly through collaborations, encounters with more data, and neuroimaging community interactions such as the one at the core of this Research Topic project. It should be emphasized again that even beyond "including" and "excluding" subjects from a study, the larger—and perhaps more important—perspective of this process is to become confident of the contents of the data being analyzed. This principle applies to both public data that has been downloaded (which may or may not have been curated, or might have been curated with different analyses in mind) as well as to locally acquired data. Scanner upgrades, manual entry to scanner consoles, "automatic" console settings (that can change due to subject weight, for example), and more can affect the properties of acquired data in subtle but important ways. The researcher always has the responsibility to be aware of the dataset contents and their relative applicability for a given study.

Quality control, in the holistic sense emphasized throughout this paper, should start at the earliest stages of a study. Researchers should be "close to their data" from the very

beginning, to reduce chances of downstream problems. Consider the following four steps:

1. Perform GTKYD, APQUANT and APQUAL checks, and review the results systematically.
2. Compare GTKYD, APQUANT and APQUAL results with previous studies.
3. (for task data) Review the duration and ISI statistics from any stimulus timing files.
4. Use the GUI to check steps of the processing (in particular running the automatically generated InstaCorr scripts) and look for any peculiarities.

When acquiring the first few subjects in a new project, it is important to perform a detailed review of the QC results across all stages, performing Steps 1–4; the same applies when starting with a shared data collection, examining a few subjects in detail. Any problems or questions should be dealt with immediately, to avoid data waste. After this in-depth review of the first few subjects, Steps 1–3 can be performed for the

remaining subjects, with GUI investigations performed if any abnormalities are found.

The QC procedure of filtering subjects from further analysis is a subtle one: A researcher must balance the goal of basing results on reliable, non-artifactual data with the need to avoid introducing a bias. To date, there are no universal set of criteria for this process, and the heterogeneity of acquisition techniques, subject populations, research questions and analysis methods suggest this would be a challenging task. For any QC criterion, a desired trait is that in practice results are not overly sensitive to its thresholding value. For example, if a small change in a quantitative threshold leads to a large change in subjects excluded, one might try to find an alternate QC measure with a better delineation. One expects that over time and with more experience and feedback, QC measures will evolve to improve fMRI analysis.

Both quantitative and qualitative criteria have unique benefits; in many cases, they provide complementary checks and verifications. Quantitative ones are easier to apply uniformly, but in fact many quantities and their threshold values are based on much qualitative “training” and experience with datasets (and the many ways in which artifactual features can arise). Qualitative criteria require particular attention to be applied consistently, but, as evinced here, they provide a necessary perspective on data that is otherwise missed due to the inherently large data compression of derived quantities. If possible, qualitative criteria should not be central to the current analysis (to avoid bias), though that may not always entirely be possible (in which case, one must rely on the consistency of assumptions).

The primary QC criteria presented here relied on derived quantities (in the GTKYD, APQUANT, and STIM stages) and static images (in APQUAL) of the data. These are useful and able to be generated in automatic and systematic ways during processing (in the present work, *via* `afni_proc.py`). However, in some cases such items may only flag *potential* data quality issues, and a full understanding requires exploring the data itself more deeply. EPI datasets are inherently 4-dimensional, and occasionally too much information has been lost within the 1-dimensional scalar quantities or 3D image montages to understand an observed feature. Interactive exploration is then necessary to avoid “false rejections” of usable data (which is wasteful and may bias results) and “false inclusions” of problematic data (which introduce non-physiological features and again may bias or distort results). Here, we showed how GUI interactions could be used to more fully explore the underlying properties of the data, particularly with AFNI’s InstaCorr⁴.

In applying these QC principles and tools to the examination of this project’s eight publicly available datasets, we found quite a number of issues that ranged from incorrect header information

(coordinate orientations, left-right flips) to ingrained data issues (temporally correlated artifacts, significant distortion, ghosting and dropout). In general, the exclusion criteria applied here were relatively light; some features such as inconsistent voxel size or acquisition parameters could be cause for rejection in an actual research study. Similarly, many datasets had distortion, dropout or other artifacts that particularly affected local brain regions, but the extent was judged as not severe enough for removal here. For a particular study’s hypotheses, though, such localized issues may render a subject’s data unusable. In the end, sizable fractions of these groups contained datasets that were categorized for exclusion, and another fraction with uncertain features for additional examination. Two groups contained systemic artifacts, likely rendering the data problematic for further analyses. This points to the necessity of performing full QC, and we hope that this Research Topic elevates QC’s role in the neuroimaging field: *Understanding* the data is an important part of processing it.

The data collections presented here provided an illustrative subset of the issues that exist in fMRI data. Many other problematic features can appear, such as major dropout from bad coils, zipper-striped artifacts, signal saturation, mechanical features in time series (e.g., from anesthesia devices), and more. Furthermore, different acquisition methods or processing choices will lead to different QC checks. For surface-based analysis, one would want to visualize the accuracy of the surface mesh estimation. For multi-echo fMRI, one might visualize maps of estimated T2*, as well as any temporal components projected out of the time series data. When combining data from several scanners, sites or even studies, the heterogeneity of datasets might prompt another layer of QC comparisons. It is important to note that QC criteria will never be set in stone, but will need to be adjusted based on the type of analysis, the subjects, and scanner and acquisition properties, which change over time.

The exact role of QC in determining final group outcomes is not well known (at present, at least). Certainly, cross-study accuracy, reproducibility and reliability should be improved by reducing artifacts in data collections. “Big data” does not preclude the need for reasonable QC—having a large fraction of problematic/artifact-heavy subjects can still be a problem whether the number of subjects is $N = 50$ or $N = 5000$. The QC process does require time and effort, but it is always a small fraction of the total effort that must be put into the study: Grant writing, pilot studies, subject recruitment, scanning, processing/reprocessing and (hopefully) publication. Choosing to save a relatively miniscule amount of QC time within a project can be quite costly, if the final results of a team’s work end up being based on unreliable data. Furthermore, if detailed QC is practiced at the early stages of data gathering, one would also expect it to greatly reduce the overall time of QC, because subtle

⁴ We note that the APQC HTML’s “vstat” block of images for resting state is essentially a quick, systematized version of InstaCorr exploration.

issues could be observed and addressed before the number of subjects grows large.

There is often a desire to reduce all QC to a simplified, automated process. However, all quantities and thresholds used in QC procedures have been based on visualizing a large number of datasets and understanding their contents in depth. Even now, our current understanding of fMRI data quality is incomplete. Moreover, this process will always evolve: Study designs vary, and the technology of data acquisition is always changing. Image and time series visualization is the key to understanding data, and this layer should not be omitted from processing and quality evaluation. Ignoring visualization reduces the strength of QC, and hinders the ability to improve and develop new QC criteria—even quantitative ones. The QC results from this current project reinforce the importance of visualization: Researchers (particularly trainees just starting in the field) need to understand the data being processed, in order to avoid basing conclusions on unreliable datasets.

Conclusion

This work addresses the question, “When should fMRI quality control be done?” with a resounding answer: “Early and often.” We present our approach to QC of fMRI data, organized as a set of stages that are integrated into standard processing with the AFNI software package. One aspect of this is evaluating subject datasets to be either included or excluded for a group level analysis. But the larger goal of the presented procedure is for researchers to deeply understand the contents of their data and to be sure of its appropriateness for their analyses of interest. This procedure applies when acquiring one’s own datasets, but remains vital when using publicly available or shared datasets. In all cases, a researcher has the responsibility to assess the properties of the data collection, and our approach here has been designed to facilitate this process with multiple layers of QC investigation. It includes a mix of scriptable, automated, visual and user-interactive checks that reinforce each other, many of which are created as standard outputs of the `afni_proc.py` pipeline generating tool. The stages begin with verifying the fundamental properties of the datasets, and continue through the single subject modeling. Using the real, public data provided in this Research Topic project, we have shown how each QC stage provided vital information about subjects for determining the suitability to include in further analyses. The range of issues present in this real data shows the continuing need for such QC procedures. We hope that researchers, data repository managers and particularly trainees in the field will find these methods and provided scripts useful when working with their own data.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found here: <https://osf.io/qaesm/wiki/home>.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Funding

This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). RR, DG, and PT were supported by the NIMH Intramural Research Program (ZICMH002888) of the NIH/HHS, USA.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2022.1073800/full#supplementary-material>

References

- Allen, E., Erhardt, E., and Calhoun, V. (2012). Data visualization in the neurosciences: Overcoming the curse of dimensionality. *Neuron* 74, 603–608. doi: 10.1016/j.neuron.2012.05.001
- Biswal, B., Mennes, M., Zuo, X., Gohel, S., Kelly, C., Smith, S., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Caballero-Gaudes, C., and Reynolds, R. (2017). Methods for cleaning the BOLD fMRI signal. *Neuroimage* 154, 128–149. doi: 10.1016/j.neuroimage.2016.12.018
- Cox, R. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173.
- Cox, R., and Glen, D. (2013). “Nonlinear warping in AFNI,” in *Proceedings of the presented at the 19th annual meeting of the organization for human brain mapping*, Seattle, WA.
- Di Martino, A., Yan, C., Li, Q., Denio, E., Castellanos, F., Alaerts, K., et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Fischl, B., and Dale, A. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11050–11055.
- Foerster, B., Tomasi, D., and Caparelli, E. (2005). Magnetic field shift due to mechanical vibration in functional magnetic resonance imaging. *Magn. Reson. Med.* 54, 1261–1267. doi: 10.1002/mrm.20695
- Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., et al. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* 54, 313–327. doi: 10.1016/j.neuroimage.2010.07.033
- Glen, D., Taylor, P., Buchsbaum, B., Cox, R., and Reynolds, R. (2020). Beware (Surprisingly common) left-right flips in your MRI Data: An efficient and robust method to check MRI dataset consistency using AFNI. *Front. Neuroinformatics* 14:18. doi: 10.3389/fninf.2020.00018
- Gohel, S., and Biswal, B. (2015). Functional integration between brain regions at rest occurs in multiple-frequency bands. *Brain Connect.* 5, 23–34. doi: 10.1089/brain.2013.0210
- Jo, H., Saad, Z., Simmons, W., Milbury, L., and Cox, R. (2010). Mapping sources of correlation in resting state FMRI, with artifact detection and removal. *Neuroimage* 52, 571–582. doi: 10.1016/j.neuroimage.2010.04.246
- Markiewicz, C., Gorgolewski, K., Feingold, F., Blair, R., Halchenko, Y., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *Elife* 10:e71774. doi: 10.7554/eLife.71774
- Saad, Z., Glen, D., Chen, G., Beauchamp, M., Desai, R., and Cox, R. (2009). A new method for improving functional-to-structural MRI alignment using local Pearson correlation. *Neuroimage* 44, 839–848. doi: 10.1016/j.neuroimage.2008.09.037
- Saad, Z. S., Reynolds, R. C., Jo, H. J., Gotts, S. J., Chen, G., Martin, A., et al. (2013). Correcting brain-wide correlation differences in resting-state FMRI. *Brain Connect.* 3, 339–352. doi: 10.1089/brain.2013.0156
- Shirer, W., Jiang, H., Price, C., Ng, B., and Greicius, M. (2015). Optimization of rs-fMRI pre-processing for enhanced signal-noise separation, test-retest reliability, and group discrimination. *Neuroimage* 117, 67–79. doi: 10.1016/j.neuroimage.2015.05.015
- Song, S., Bokkers, R., Edwardson, M., Brown, T., Shah, S., Cox, R., et al. (2017). Temporal similarity perfusion mapping: A standardized and model-free method for detecting perfusion deficits in stroke. *PLoS One* 12:e0185552. doi: 10.1371/journal.pone.0185552
- Taylor, P., Reynolds, R., Calhoun, V., Gonzalez-Castillo, J., Handwerker, D., Bandettini, P., et al. (2022). Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. *bioRxiv* [Preprint]. doi: 10.1101/2022.10.26.513929



OPEN ACCESS

EDITED BY

Daniel R. Glen,
National Institute of Mental Health (NIH),
United States

REVIEWED BY

Adrian W. Gilmore,
National Institute of Mental Health (NIH),
United States
Andrew Jahn,
University of Michigan, United States

*CORRESPONDENCE

Brendan Williams
✉ b.williams3@reading.ac.uk

†These authors have contributed equally to this work

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

RECEIVED 14 October 2022

ACCEPTED 11 January 2023

PUBLISHED 03 February 2023

CITATION

Williams B, Hedger N, McNabb CB,
Rossetti GMK and Christakou A (2023)
Inter-rater reliability of functional MRI data
quality control assessments: A standardised
protocol and practical guide using pyfMRIqc.
Front. Neurosci. 17:1070413.
doi: 10.3389/fnins.2023.1070413

COPYRIGHT

© 2023 Williams, Hedger, McNabb, Rossetti and
Christakou. This is an open-access article
distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with
these terms.

Inter-rater reliability of functional MRI data quality control assessments: A standardised protocol and practical guide using pyfMRIqc

Brendan Williams^{1,2*}, Nicholas Hedger^{1,2†}, Carolyn B. McNabb^{3†},
Gabriella M. K. Rossetti^{1,2†} and Anastasia Christakou^{1,2}

¹Centre for Integrative Neuroscience and Neurodynamics, University of Reading, Reading, United Kingdom,

²School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom,

³Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, College of Biomedical and Life Sciences, Cardiff University, Cardiff, United Kingdom

Quality control is a critical step in the processing and analysis of functional magnetic resonance imaging data. Its purpose is to remove problematic data that could otherwise lead to downstream errors in the analysis and reporting of results. The manual inspection of data can be a laborious and error-prone process that is susceptible to human error. The development of automated tools aims to mitigate these issues. One such tool is pyfMRIqc, which we previously developed as a user-friendly method for assessing data quality. Yet, these methods still generate output that requires subjective interpretations about whether the quality of a given dataset meets an acceptable standard for further analysis. Here we present a quality control protocol using pyfMRIqc and assess the inter-rater reliability of four independent raters using this protocol for data from the fMRI Open QC project (<https://osf.io/qaesm/>). Data were classified by raters as either “include,” “uncertain,” or “exclude.” There was moderate to substantial agreement between raters for “include” and “exclude,” but little to no agreement for “uncertain.” In most cases only a single rater used the “uncertain” classification for a given participant’s data, with the remaining raters showing agreement for “include”/“exclude” decisions in all but one case. We suggest several approaches to increase rater agreement and reduce disagreement for “uncertain” cases, aiding classification consistency.

KEYWORDS

fMRI, resting state fMRI, task fMRI, quality control, inter-rater reliability

Introduction

Functional magnetic resonance imaging (fMRI) data are inherently multi-dimensional with many potential sources of artefacts that can lead to spurious results (Power et al., 2012; Van Dijk et al., 2012). Therefore, ensuring data are of sufficient quality for analysis is an essential step in the processing of fMRI data. This is especially important for large multi-site studies such as the Adolescent Brain Cognitive Development study (Casey et al., 2018), and the Human Connectome Project (Van Essen et al., 2013), where time required to perform detailed, manual screening of individual data can quickly become intractable. To address this, many quality control tools and pipelines now exist to help users make informed decisions about quality in

their datasets (Marcus et al., 2013; Esteban et al., 2017; Alfaro-Almagro et al., 2018). These tools—which automate part of the quality control process—aim to decrease the time taken to assess data quality, minimise the amount of prior knowledge needed to make informed decisions, and reduce errors during assessment.

Several tools currently exist for assessing the quality of fMRI data, including MRIQC (Esteban et al., 2017), and Visual QC (Raamana, 2018). This list also includes pyfMRIqc, which we developed at the Centre for Integrative Neuroscience and Neurodynamics (CINN), University of Reading (Williams and Lindner, 2020). Many of our neuroimaging facility users at CINN are Ph.D. students and early career researchers, who join a community that prioritises practical training and learning opportunities. As part of this commitment, we develop software which is user-friendly and empowers individuals to become confident and informed researchers. pyfMRIqc helps users to make informed decisions about the quality of their data by generating various image quality metrics and presenting them in an easily interpretable way in a visual report. pyfMRIqc also has extensive online documentation that describes to users how these plots are generated and what they show, and aids their interpretation with examples. Users of pyfMRIqc can generate these reports with minimal programming experience, requiring only a single line of code to run the software and without the need for using containerised environments for generating output. As part of the work presented here, we additionally developed a piece of software, “cinnqc,” which we used to automate the minimal pre-processing and curation of data for pyfMRIqc, and to identify cases where data deviate from the expected acquisition parameters for the dataset.

Previous reports describe the use of inter-rater reliability for the quality assessment of structural imaging data (Backhausen et al., 2016; Esteban et al., 2017; Rosen et al., 2018; Benhajali et al., 2020). For instance, Benhajali et al. (2020) developed a method for quickly assessing the registration of T1 weighted images to standard MNI space. Raters included citizen scientists who had no previous experience with MRI data, as well as expert raters. Their protocol resulted in good reliability, particularly with respect to which images were deemed to fail quality assessment, between expert raters, with citizen scientists also showing agreement. The study therefore demonstrated that this straightforward approach for assessing registration quality was consistent between individuals with different skill levels. Another protocol assessed for reliability between raters was presented by Backhausen et al. (2016), who aimed to provide a workflow for the quality control assessment of T1 images both during and after image acquisition to maximise useful sample size. Images were classified into three categories (pass, check, fail), and these three categories were associated with significant differences in cerebral cortex, left amygdala, and total grey matter volume estimations. Reliability between two raters for the three classification categories was high [intra-class correlation coefficient ($\alpha = 0.931$)], in line with results from Rosen et al. (2018), who found good consistency between expert raters when a three category rating system was used (although notably concordance was significantly lower when using five categories). Lastly, Esteban et al. (2017) demonstrated fair to moderate agreement between two raters when assessing the quality of T1 data from the ABIDE dataset. These studies demonstrate that reasonable reliability can be expected of subjective decisions about the quality of structural imaging data, particularly when three categories are used to classify data. However, in the case of functional data, and despite its potential utility, inter-rater reliability has not been similarly evaluated to help understand the consistency of subjective decisions about data quality. To asse-

whether experienced raters are reliable in their classifications of functional data quality across datasets, we used data from the fMRI Open QC project,¹ which included data with different acquisition parameters from multiple sites.

We assess the inter-rater reliability of fMRI data quality assessments for task-based and resting state data. We describe quantitative and qualitative criteria for classifying data quality, present a quality control protocol for assessing raw fMRI data quality using pyfMRIqc, assess reliability between four independent raters using this protocol, and provide example cases of different data quality issues using output from pyfMRIqc. Raters classified data into one of three assessment categories, “include,” “uncertain,” or “exclude.” Using our protocol, we find moderate to substantial reliability between raters, particularly for “include” and “exclude” decisions, but less agreement between raters for the uncertain classification.

Materials and methods

Participants

Imaging data participants

Imaging data from 129 subjects were included. Each subject had a T1 weighted high-resolution anatomical image, and a single-band echo-planar imaging (EPI) image for either task-based or resting state functional magnetic resonance imaging (fMRI) acquisition. Task-based fMRI data were included for 30 subjects. Resting-state fMRI data were included for 99 subjects; resting-state data originated from five sites, with approximately 20 subjects per site. Data originated from the following publicly available datasets: ABIDE, ABIDE-II, Functional Connectome Project, and OpenNeuro (Biswal et al., 2010; Di Martino et al., 2014; Markiewicz et al., 2021). Data from each site were treated as separate datasets for the purpose of performing quality assessment. The expected acquisition parameters for data from each site are summarised in Table 1. The data presented here are available on the Open Science Framework page of the fMRI Open QC project (see text footnote 1).

Quality control raters

Quality control assessments were completed by four independent raters (BW, NH, CBM, GMKR), who were all postdoctoral research fellows, and all raters had previous experience in quality assessment, processing and analysis of functional neuroimaging data. Two raters (BW and GR) had previously used pyfMRIqc to perform quality assessment of fMRI data. Additionally, BW was involved in the development of pyfMRIqc. Each rater reviewed data for 104 of the 129 subjects, using outputs from cinnqc and pyfMRIqc. Subject assignment ensured at least four subjects from each site were reviewed by all four raters, and every other subject was reviewed by three raters. Assignments were also balanced so that the proportion of overlapping cases was equal across raters (see Supplementary Data Sheet 1 for details of rater assignments).

Data processing

Minimal pre-processing of anatomical T1 weighted and functional EPI data was performed using the FSL toolbox (version

¹ <https://osf.io/qaesm/>

TABLE 1 Expected acquisition parameters for subjects in each site in the main dataset.

Subjects	Modality	Voxel size (mm)	Matrix	Volumes	TR (s)
sub-001 → sub-030	T1w	1 × 1 × 1	176 × 256 × 256	1	
	EPI	3 × 3 × 4	64 × 64 × 34	242	2
sub-101 → sub-120	T1w	1 × 1 × 1	256 × 200 × 256	1	
	EPI	2.67 × 2.67 × 3	96 × 96 × 47	156	2.5
sub-201 → sub-220	T1w	1 × 1 × 1	160 × 256 × 256	1	
	EPI	3 × 3 × 3.840789	80 × 80 × 38	150	2
sub-301 → sub-316	T1w	0.976562 × 1.2 × 0.976562	256 × 182 × 256	1	
	EPI	1.5625 × 1.5625 × 3.1	128 × 128 × 45	162	2.5
sub-401 → sub-423	T1w	1 × 1 × 1	256 × 200 × 256	1	
	EPI	2.667 × 2.667 × 3	96 × 96 × 47	123	2.5
sub-701 → sub-720	T1w	1 × 1 × 1	192 × 256 × 256	1	
	EPI	3 × 3 × 3.51	64 × 64 × 39	198	2.5

T1w modality is the high-resolution T1 weighted anatomical image. EPI modality is the functional (BOLD) task-based (sub-001 → sub-030) and resting state (sub-101 → sub-720) echo-planar images. TR is the time taken in seconds to acquire a single volume of EPI data.

6.0) from the Oxford Centre for Functional MRI of the Brain (FMRIB's Software Library²) (Jenkinson et al., 2012). Data pre-processing, curation, and quality control was automated using “cinnqc.”³ cinnqc provides wrapper scripts for executing and curating output from FSL pre-processing functions (e.g., motion correction, registration, and brain extraction), and also generating pyfMRIqc reports for minimally pre-processed data. To pre-process data, the T1 image was skull stripped using the Brain Extraction Tool (Smith, 2002), then grey matter, white matter, and cerebrospinal fluid tissue segmentation was performed using FMRIB's Automated Segmentation Tool (Zhang et al., 2001). Functional EPI data were motion corrected with MCFLIRT (Jenkinson et al., 2002), using affine transformations to align the first volume of functional data with each subsequent volume. Functional EPI and anatomical T1 data were then co-registered using the epi_reg function,⁴ and a linear affine transformation was used to convert a brain extracted mask of the T1 anatomical image to functional EPI space using FMRIB's Linear Image Registration Tool (Jenkinson and Smith, 2001; Jenkinson et al., 2002). The brain mask in functional EPI space was then re-binarised using a threshold of 0.5. Image quality metrics and plots were generated using pyfMRIqc (Williams and Lindner, 2020) to aid data quality assessment, e.g., the identification of artefacts that were participant-, sequence-, technique-, or tissue-specific. pyfMRIqc was run with the following input arguments: -n < motion corrected EPI data >, -s 25, -k < brain extracted mask in functional space > -m < motion parameter output from MCFLIRT >.

Resources

Ubuntu 20.04.4 LTS

FSL version 6.0 (see text footnote 2).

Anaconda 4.10.1

Python 3.8.8

- cinnqc 0.1.0
- easygui 0.98.3
- matplotlib 3.3.4
- nibabel 3.2.1
- numpy 1.20.1
- pandas 1.2.4

Quality assessment protocol

Raters were given the following instructions before beginning quality assessment:

TABLE 2 Quantitative criteria for determining dataset inclusion/exclusion.

Criteria	Exclusion criteria
Motion	Any relative movements > Voxel size More than 5 relative movements > 0.5 mm ¹ Max absolute motion > 2 mm (1.5 mm is marginal) ¹
Slice-wise SNR	< 99 (99 → 150 is marginal) ^{1*}
Consistent voxel sizes	No ² (to 2d.p.)
Consistent number of volumes	No ²
Consistent number of scans in the dataset	No ³
T1w whole brain coverage	No ⁴
EPI whole brain coverage in the mean image of the pyfMRIqc report and the first volume	No ⁴

*Some slices will return slice-wise TSNR values of NaN. NaN values are returned because the slice does not have any voxels that SNR are calculated for; if this is the case, then the presence of these NaN values should not be used for the purpose of exclusion. Some slices will include a large proportion of non-brain voxels which will have lower values relative to brain voxels decreasing the slice-wise TSNR mean. If this is the case then use your discretion in your assessment of slice-wise TSNR.

¹: Center for Brain Science, Harvard University (<https://cbs.fas.harvard.edu/facilities/neuroimaging/investigators/mr-data-quality-control/>); ²: Human Connectome Project (Marcus et al., 2013); ³: BIDS standard (Gorgolewski et al., 2016); ⁴: UK Biobank (Alfaro-Almagro et al., 2018).

2 www.fmrib.ox.ac.uk/fsl

3 <https://github.com/bwilliams96/cinnqc>

4 https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT/UserGuide#epi_reg

The following criteria need to be used to classify all images⁵ :

- Include—no quality assessment issues that indicate the dataset is problematic.
- Uncertain—some quality assessment issues that makes the inclusion of dataset marginal.
- Exclude—quality assessment issues that mean the data should not be included.

Each image classified as either “uncertain” or “exclude” should include an explanation of why the given classification was made. Please be as descriptive as possible when explaining your decision-making.

Quality assessment decision-making should be supported by the output produced by *cinnqc* and *pyfMRIqc*. *cinnqc* and *pyfMRIqc* derivatives can be found online in the directories `/cinnqc/examples/{fmriqc-open-qc-task, fmriqc-open-qc-rest-100, fmriqc-open-qc-rest-200, fmriqc-open-qc-rest-300, fmriqc-open-qc-rest-400, fmriqc-open-qc-rest-500, fmriqc-open-qc-rest-600, fmriqc-open-qc-rest-700}/derivatives/cinnqc/` of the *cinnqc* GitHub page (see text footnote 3).

Quantitative data assessment

Quantitative quality assessment criteria for T1 and EPI data based on acquisition parameters and derived metrics from the data are summarised in [Table 1](#). Thresholds for absolute and relative motion, as calculated using MCFLIRT, are given to limit its effect on data quality. Motion thresholds are defined in [Table 2](#) and are summarised in the *pyfMRIqc* report. Yet, even motion that is sub-threshold could still impact data quality. Qualitative data assessment should be carried out to check whether any motion incidents coincide with a problematic change in signal. Temporal signal to noise (TSNR, referred to as SNR in *pyfMRIqc*) is calculated as mean intensity divided by the standard deviation of voxels (25th centile mean intensity) outside the brain-extracted mask in functional space. It is calculated by *pyfMRIqc* on minimally pre-processed data. Slice-wise TSNR should be checked in the *pyfMRIqc* report, and potentially problematic slices should be followed up using qualitative assessments. Field of view, number of volumes, and scans are checked using *cinnqc*, and a file with the suffix `*_notes.txt` is generated to describe any potential issues. Note, some voxel dimensions may appear to be different due to rounding, but if they are equal to 2 decimal places then subjects do not need to be excluded. T1 and EPI data should have whole brain coverage, which includes the cerebral cortex and subcortical brain regions (but not necessarily the cerebellum). A summary of quantitative assessment criteria can

be found in [Table 2](#), and a summary of the expected acquisition parameters can be found in [Table 1](#).

Qualitative data assessment

pyfMRIqc generates a number of plots and tables that can be helpful in the qualitative assessment of data. Mean and slice-wise scaled squared difference (SSD) is calculated by squaring the difference in voxel intensity between consecutive volumes, and dividing by the global mean squared difference. In the QC plots section, mean and slice-wise SSD graphs can be used to identify global, and slice-wise changes in signal intensity, respectively. SSD is also plotted alongside the global normalised mean voxel intensity, normalised SSD variance, plus absolute and relative motion to visualise relationships between changes in SSD, signal intensity, and motion. Further, mean, minimum, and maximum SSD is plotted slice-wise to determine whether issues are present in specific slices.

The plot of the “Mean voxel time course of bins with equal number of voxels” is generated by binning voxels into 50 groups, based on their mean intensity, and calculating the mean intensity for voxels in each bin for each volume. Bins are ordered top-down from lowest mean intensity voxels (non-brain/cerebrospinal fluid) to highest (grey matter, then white matter voxels). This plot enables easy visualisation of signal variance and was originally described by [Power \(2017\)](#), where further information can also be found.

The “Masks” plot can be helpful in indicating whether there were issues during acquisition or processing (such as brain extraction and/or registration of T1 and EPI data). For instance, there may be many brain voxels that are not highlighted in blue. If this is the case, then scans should be carefully checked for signal distortion (described below), or processing steps may need to be manually re-run with adjusted input parameters. Poor registration (for instance, misalignment of gross anatomical structures including brain surface, or grey matter/white matter/cerebrospinal fluid boundaries) may be indicative of other data quality issues.

The “Variance of voxel intensity” plot visualises the variance in signal in each voxel over the timeseries of the functional run. The png image given in the *pyfMRIqc* report is thresholded (voxel intensities are divided into 1,000 equal width bins, and the intensity of the highest bin with at least 400 voxels is used) to aid visualisation, however a nifti version of the image is also included which is unthresholded. This nifti image is useful for more in-depth investigation if there are potential quality issues or the figure appears problematic. The “Sum of squared scaled difference over time” plot presents the voxel-wise sum of SSD over the functional run. Similarly to the “Variance of voxels intensity” plot, we applied a threshold for the png figure for readability (sum of squared scaled differences are divided into 50 equal width bins, and the upper threshold of the fifth bin is used), but the nifti image does not have a threshold.

To inspect data for signal distortion, load T1 images from the subject’s BIDS directory; for EPI images, load the image with the suffix `*_example-func.nii.gz` from the subject’s *cinnqc* BIDS derivative directory, and the mean voxel intensity nifti file from *pyfMRIqc*. If visual abnormalities are present, this could impact the signal (e.g., image distortion, signal loss, artefacts such as ringing or ghosting), or processing (e.g., brain extraction, registration, motion correction) of T1 or EPI data. To determine if this is the case, the plots from *pyfMRIqc* can be used to aid subject classification. Detailed explanations for interpreting *pyfMRIqc* plots and tables can be found

⁵ Additional information about classifications not given in the protocol but which was agreed by raters:

- Include cases would pass all quantitative and qualitative quality control criteria and *pyfMRIqc* plots or manual inspection of data would not indicate any issues with data.
- Uncertain cases would pass all quantitative quality control criteria, but *pyfMRIqc* plots or manual inspection of data may indicate marginal issues in the data that could warrant exclusion.
- Exclude cases would fail at least one quantitative quality control criteria, and/or *pyfMRIqc* plots or manual inspection of data indicate data quality issues that would warrant exclusion.

TABLE 3 Qualitative criteria for determining dataset inclusion.

Criteria	Threshold
Aberrant pyfMRIqc output	Plots or tables that indicate problematic EPI data, supported by visual inspection of functional data
T1w signal distortion	Visual abnormalities in the acquisition of the T1w image, such as ringing artefacts that would impair registration to standard template
EPI signal distortion	Visual abnormalities in the mean image of the pyfMRIqc report or the first volume of the fMRI data that would impair registration to standard template
Atypical brain structure	Morphology that would impair registration to standard template (pathological or non-pathological).

TABLE 4 Example cases that were used during the calibration session for raters before independently assessing the whole dataset.

Subject	Include	Uncertain	Exclude	Notes
sub-013			1	Many volumes with relative movement > 0.1. Motion events around volumes 65 and 205 appear to cause global decrease in signal
sub-103	1			Peak in SSD between volumes 95–100 looks like its driven by eye movement
sub-207			1	More than 5 relative motion events > 0.5. Max absolute movement is marginal

in the pyfMRIqc User Manual.⁶ A summary of qualitative assessment criteria can be found in Table 3.

Rater calibration and reliability assessment

Each rater independently assessed and classified subjects using the quality assessment protocol described above. To ensure quality assessment criteria were interpreted consistently, BW used the quality assessment protocol to identify exemplar subjects for issues and presented these training cases to the other raters (Table 4).

Fleiss' kappa (Fleiss, 1971) was calculated using the “irr” package in R (Gamer et al., 2019) to assess pair-wise and category-wise inter-rater reliability between raters; to correct for multiple comparisons we used the Holm method to control the family-wise error rate using the “p.adjust” function in R (Holm, 1979; R Core Team, 2020). We chose to use Fleiss' kappa instead of Cohen's kappa, because Fleiss' kappa also allows us to determine how similar pairs of raters are across classifications by calculating category-wise agreement.

We used the criteria described by Landis and Koch (1977) to interpret Fleiss' kappa using the following benchmarks to describe the strength of agreement: poor agreement < 0.00; slight agreement 0.00–0.20; fair agreement 0.21–0.40; moderate agreement 0.41–0.60; substantial agreement 0.61–0.8; almost perfect agreement 0.81–1.00. Overall agreement across raters and categories was calculated using Krippendorff's alpha (Krippendorff, 1970), which is useful as a measure of overall agreement because it is not restricted by the number of raters or the presence of missing data in the sample (Hayes and Krippendorff, 2007). Krippendorff's alpha and bootstrap 95% confidence intervals (1,000 iterations, sampling subjects with replacement) were calculated in R using scripts from Zapf et al. (2016).

Results

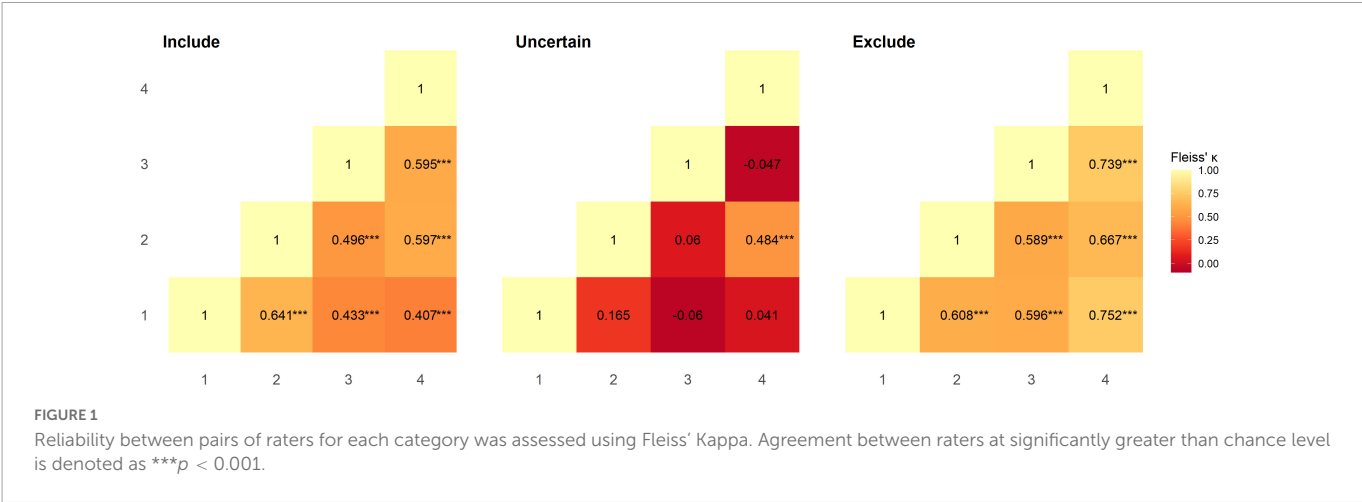
Each subject was categorised as either “include” (rater one: 68, rater two: 73, rater three: 80, rater four: 74), “uncertain” (rater one: 10, rater two: 12, rater three: 3, rater four: 9), or “exclude” (rater one: 26, rater two: 19, rater three: 21, rater four: 20) by the four raters. Overall percentage agreement between raters is summarised in Table 5.

Inter-rater reliability between pairs of raters was calculated using Fleiss' Kappa; overall agreement between all pairs of raters was moderate and significantly greater than chance level (rater 1–2: $\kappa = 0.536$, $z = 6.143$, $p < 0.001$; rater 1–3: $\kappa = 0.437$, $z = 4.639$, $p < 0.001$; rater 1–4: $\kappa = 0.456$, $z = 5.17$, $p < 0.001$; rater 2–3: $\kappa = 0.448$, $z = 5.071$, $p < 0.001$; rater 2–4: $\kappa = 0.596$, $z = 6.818$, $p < 0.001$; rater 3–4: $\kappa = 0.578$, $z = 6.022$, $p < 0.001$). Category-wise Kappa for all raters was moderate and substantial for “include” and “exclude” assignments respectively and was significantly greater than chance level (“include”: $\kappa = 0.514$, $z = 6.661$, $p < 0.001$; “exclude”: $\kappa = 0.731$, $z = 9.472$, $p < 0.001$). However, this was not the case for “uncertain” assignments, where agreement between raters was slight ($\kappa = 0.013$, $z = 0.166$, $p = 1.0$). We also calculated Fleiss' Kappa category-wise for pairs of raters (Figure 1). All raters had moderate to substantial agreement, and performed at significantly greater than chance level for “include” (rater 1–2: $z = 5.697$, $p < 0.001$; rater 1–3: $z = 3.849$, $p < 0.001$; rater 1–4: $z = 3.591$, $p < 0.001$; rater 2–3: $z = 4.409$, $p < 0.001$; rater 2–4: $z = 5.269$, $p < 0.001$; rater 3–4: $z = 5.253$, $p < 0.001$) and “exclude” (rater 1–2: $z = 5.405$, $p < 0.001$; rater 1–3: $z = 5.3$, $p < 0.001$; rater 1–4: $z = 6.645$, $p < 0.001$; rater 2–3: $z = 5.231$, $p < 0.001$; rater 2–4: $z = 5.887$, $p < 0.001$; rater 3–4: $z = 6.525$, $p < 0.001$) assignments, but not for “uncertain” (rater 1–2: $z = 1.471$, $p = 1.0$; rater 1–3: $z = -0.537$, $p = 1.0$; rater 1–4: $z = 1.0$, $p = 0.716$; rater 2–3: $z = 0.529$, $p = 1.0$; rater 2–4: $z = 4.272$, $p < 0.001$; rater 3–4: $z = -0.415$, $p = 1.0$) assignments (Figure 1). Overall agreement in the dataset, as assessed using Krippendorff's alpha was 0.508 [95% bootstrap confidence intervals (0.381, 0.615)]; removing instances where “uncertain” was assigned increased Krippendorff's

TABLE 5 Overall percentage agreement between raters for “include”/“uncertain”/“exclude” assignments.

	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	–	78.481	75.949	75.641
Rater 2	78.481	–	75.949	83.333
Rater 3	75.949	75.949	–	84.615
Rater 4	75.641	83.333	84.615	–

⁶ <https://drmichaellindner.github.io/pyfMRIqc/>



alpha to 0.694 [95% bootstrap confidence intervals (0.559, 0.802)]. In total, at least two raters categorised 97 subjects as include, 6 subjects as uncertain, and 26 subjects as exclude. **Supplementary Table 1** summarises the subject-wise group majority classification (“include”/“uncertain”/“exclude”).

QC “exclude” criteria examples

Data acquisition artefacts

Imaging acquisition artefacts were identified in five subjects by at least one rater. These issues included ghosting (aliasing), ringing, and wraparound artefacts (Table 6). In Figure 2 we present these three artefacts, with the relevant output from pyfMRIqc used to identify the issue. For the first subject, ghosting (aliasing) in the mean functional image from pyfMRIqc was detected (Heiland, 2008). This can be detected visually as the presence of spurious signal outside

the perimeter of the head. In the second case, wraparound of the functional signal was detected in the mean functional image (Arena et al., 1995). Wraparound can be detected when part of the head is partially occluded by the field of view. In this case, the most posterior portion of the head appears instead in the anterior portion of the image and is most noticeable visually on axial and sagittal slices. The third case contained in-plane artefacts in the data due to eye movements (McNabb et al., 2020). In this case, both the variance in voxel intensity, plus a peak in the maximum and sum of the scaled squared difference in affected slices (particularly slices 15–17) indicated the presence of physiologically unrelated changes in signal. These effects are especially pronounced around volumes 19–21, where there is a peak in the variance of the sum of squared difference. A video of flickering in affected slices is included in Supplementary Video 1.

Motion

30 datasets were classified as “exclude” by at least one rater with issues relating to motion described in the notes (Table 6). Of these cases, 17 exceeded acceptable values set out in our quantitative criteria for absolute and relative motion (Table 2). The remaining cases were classified as “exclude” based on the residual effects of motion upon the data, despite the quantitative measure of motion being sub-threshold (Friston et al., 1996). This includes decreases in global signal coinciding with the onset of motion events (Figure 4), plus peaks in scaled squared difference and banding in the binned carpet plot (Figure 3; Power, 2017).

Signal loss

Sudden changes in global signal can be assessed in several ways using pyfMRIqc. For instance, Figure 4 demonstrates when motion artefacts lead to a sudden decrease in global signal (Power et al., 2017). The onset of head motion around volumes 12 and 70, identified by the peaks in the mean and variance of the scaled squared difference plus the sum of relative and absolute movements, is immediately followed by a decrease in the normalised mean voxel intensity of around two standard deviations for approximately ten volumes (Figure 4). Banding is also present in the binned carpet plot, where sudden changes in signal coincide with changes in intensity across all bins. Six of the reviewed subjects were reported as having SNR related issues by at least one of the four raters, and eleven were reported as having global signal issues (Table 6).

TABLE 6 Number of datasets excluded by single, majority, and all raters for each of the relevant exclusion reasons.

	Single rater	Majority raters	All raters
Abnormal brain morphology	2	0	0
Aliasing	1	2	0
Global signal	10	1	0
Incorrect acquisition parameters	4	0	0
Motion	9	4	17
Non-whole brain coverage	1	0	0
Ringing artefact	1	0	0
SNR	5	1	0
Unidentified artefact	4	0	0
Wraparound artefact	1	0	0

In all cases where raters excluded a subject, the rater also provided notes explaining their reasons for exclusion. Here, these notes are categorised into groups, and the number of times a single, majority (two of three when three raters were assigned or two/three when four raters were assigned) or all raters mentioned that category in their notes is reported. Note that raters could give multiple reasons for excluding a subject, which means that agreement for exclusion could be based on different reasons. Subject and category-wise frequencies for single, majority, and all raters, as well as the number of raters excluding each subject are included in Supplementary Data Sheet 2.

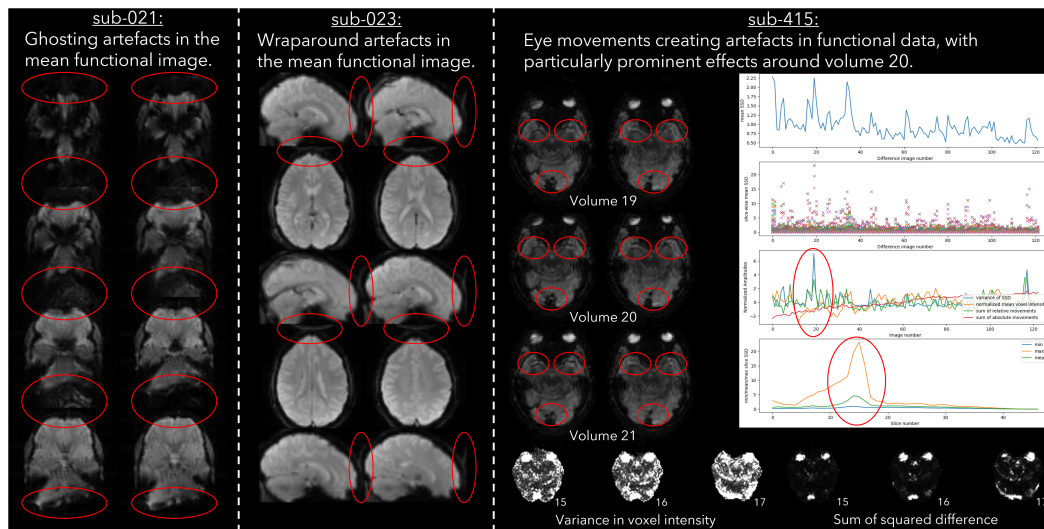


FIGURE 2
Example cases of three different types of data acquisition artefacts detected using output from pyfMRIqc.

Atypical brain structure

Two subjects were excluded by one rater due to the presence of atypical brain structure in the T1 weighted anatomical image (Table 6). Both cases are detailed in Figure 5, with one subject having a right ventricle that was enlarged and covering greater than both the extent of the left ventricle and where we would typically expect the ventricle to cover. The second subject had an unexpected mass in their left ventricle, and hypointensities in white matter across the whole brain. We are unable to comment on the clinical relevance of these anatomical features as none of the authors have clinical expertise.

Uncertain cases

27 subjects were classified as “uncertain” by at least one rater; “uncertain” was used as a classification by a single rater for 21 subjects, and more than one rater for six subjects. For the subjects classified as “uncertain” by one rater, the other two/three raters gave the same classification (“include”/“exclude”) for 20 of the 21 subjects; one subject received one “include,” one “uncertain,” and one “exclude” classification. For the remaining six cases where more than one rater classified subjects as “uncertain,” the notes for four subjects indicated the presence of issues related to residual motion that were below our threshold, while the notes for the other two subjects indicated the presence of possible pathology in the T1 image and aliasing in functional data. Lastly, only one dataset was rated as “uncertain” by all raters (Figure 4), with raters “uncertain” about the effects of sub-threshold residual motion on the data.

Discussion

This work aimed to describe a protocol for assessing the quality of raw task-based and resting state fMRI data using pyfMRIqc, and to assess the reliability of independent raters using this protocol to classify data with respect to whether it meets an acceptable standard for further analysis. We used data from the fMP

Open QC Project [(see text footnote 1), data were derived from ABIDE, ABIDE-II, Functional Connectome Project, and OpenNeuro (Biswal et al., 2010; Di Martino et al., 2014; Markiewicz et al., 2021)]. Overall, we found moderate agreement between raters, and moderate to substantial category-wise agreement between raters for include/exclude classifications. Poor to moderate category-wise agreement was found for the uncertain classification, with reliability at significantly greater than chance level for only one pair of raters. Krippendorff’s alpha for the include/exclude categories across all raters was sufficient to tentatively accept the raters’ classifications were reliable (Krippendorff, 2004, p. 241). We also provide examples for different types of quality issues that were identified in the dataset.

For the “uncertain” classification we found that there was a lack of reliability between raters, with two pairs of raters having negative κ values, indicating no agreement (McHugh, 2012), and a further three pairs having coefficients close to 0. The lack of reliability between raters for the “uncertain” classification appears to be driven by the uncertainty of a single rater for a given subject. Of the 27 subjects rated “uncertain” by any rater, 21 (78%) were not rated “uncertain” by the other raters. Of the 6 subjects rated “uncertain” by more than one rater, uncertainty related to concerns about motion ($N = 4$), aliasing ($N = 2$), and possible pathology ($N = 2$). This included one subject (sub-013) who was classified by all raters as “uncertain” due to residual effects of motion, yet other similar subjects (e.g., sub-010 and sub-016) were unanimously classified as “exclude” despite having visually similar plots, and less maximum absolute motion (Figure 3). In our quantitative exclusion criteria (Table 2), we give explicit thresholds for absolute and relative movement events, and though 17/30 excluded data sets were excluded by at least one rater due to exceeding our quantitative movement thresholds, 13/30 were excluded based on qualitative assessment of movement effects on data quality. These thresholds are relatively arbitrary, and despite being a helpful heuristic, they did not appear to capture all cases where motion had an adverse effect on data. pyfMRIqc counts the number of relative motion events $>$ voxel size, 0.5 mm and 0.1 mm, and though we set our thresholds for the number of relative motion

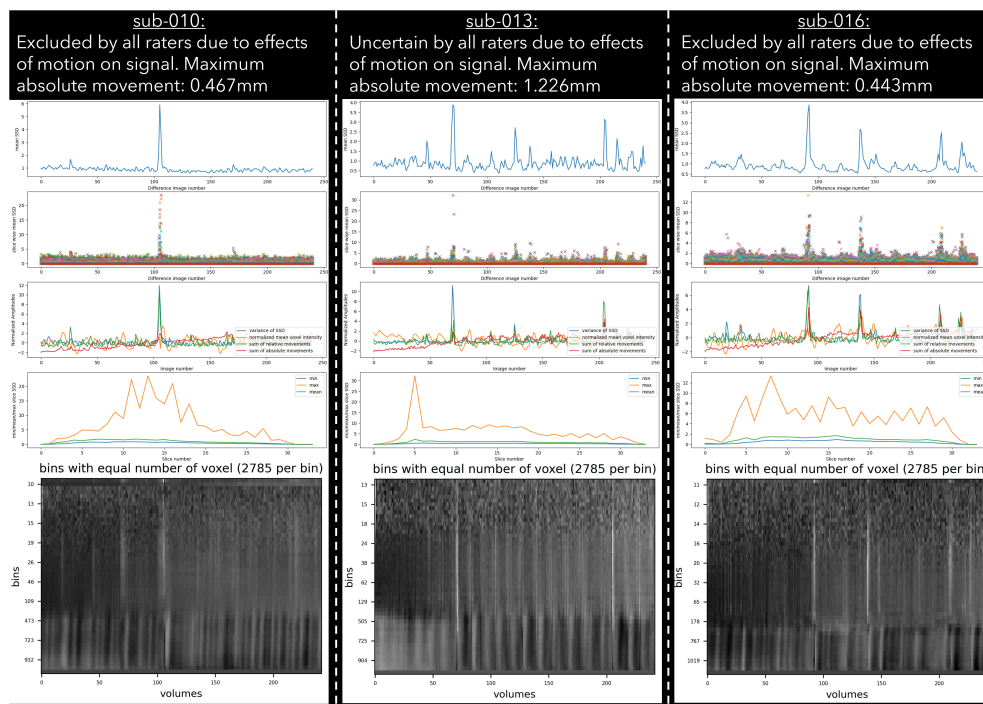


FIGURE 3

Example cases of three different types of motion artefacts detected using output from pyfMRIqc. One subject (sub-013) was classified as uncertain by all raters, while sub-010 and sub-016 were classified as exclude by all raters.

events > voxel size and 0.5 mm based on previous guidelines,⁷ we did not set a threshold for motion events > 0.1, but < 0.5 mm. In cases where motion was sub-threshold but still an issue, persistent but small motion events could negatively impact data as we did not include a threshold for small ($0.1 < \text{motion} < 0.5$) motion events. Nevertheless, it is worth mentioning that data reviewed here by raters was only minimally pre-processed and that approaches such as ICA-based denoising (Pruim et al., 2015), the inclusion of motion parameters in a model (Friston et al., 1996), and removing volumes affected by motion (Power et al., 2012) can, and often are used during data pre-processing to decrease the negative effects of motion on the signal in fMRI data. However, though these approaches are helpful for cleaning data that may otherwise be discarded, we feel that consensus guidelines for (un)acceptable levels of motion are needed to improve consistency within the neuroimaging community, in the same way the BIDS standard (Gorgolewski et al., 2016) has been widely adopted as the *de facto* data formatting structure.

It is important to note that despite the data only being minimally pre-processed, the purpose of pyfMRIqc is not to determine whether data processing steps worked as expected, but to assess the quality of the data itself. We motion corrected data so that our metrics (e.g., scaled squared difference) are calculated for contiguous voxels in time and space but we do not directly measure whether all physical motion was corrected for. Brain extraction, spatial normalisation, distortion correction, and denoising, are all commonly used and important pre-processing steps in the pipeline of fMRI data analysis, and the efficacy of these pre-processing steps should also be checked as part of a robust analysis pipeline for ensuring data quality. Therefore, the

output generated by pyfMRIqc should be treated as one part of a broader data processing procedure. Additionally, because the image quality metrics generated by pyfMRIqc have no absolute reference – that is they cannot be compared to a reference value since there is no ground truth – the detection of data quality issues is dependent on individual interpretation. One way to address this issue is by generating a database of reference values to aid outlier detection. This is the process used by MRIQC, which crowdsources image quality metrics to generate population-level distributions (Esteban et al., 2017). However, we are currently unable to generate these distributions with pyfMRIqc.

Cognitive biases may also influence subjective decision-making about the quality of fMRI data. The acquisition and preparation of an fMRI dataset involves great economic and time cost, and researchers may be motivated more by these sunk costs to minimise loss from their own data than from secondary datasets. People tend to be loss averse (Kahneman and Tversky, 1979), and the thought of “wasting” the resources put into acquiring the dataset could bias individuals to perceive data quality issues as less problematic than if the data were collected independently. For instance, Polman (2012) found that people are less loss averse when making decisions for others compared to themselves, and that this reduction of loss aversion may be due to a decreasing effect of cognitive bias on decision making. Compared with others, people also disproportionately value things they have created themselves (Norton et al., 2012), and may therefore be reluctant to discard data they perceive as having value. Reappraisal is one strategy that can be used to decrease loss aversion (Sokol-Hessner et al., 2009, 2013), and could improve decision-making by changing the perspective of discarding data from a waste of spent resource to a way of maximising ability to detect effects and improve data quality. The adoption of open research practices, such as the

⁷ <https://cbs.fas.harvard.edu/facilities/neuroimaging/investigators/mr-data-quality-control/>

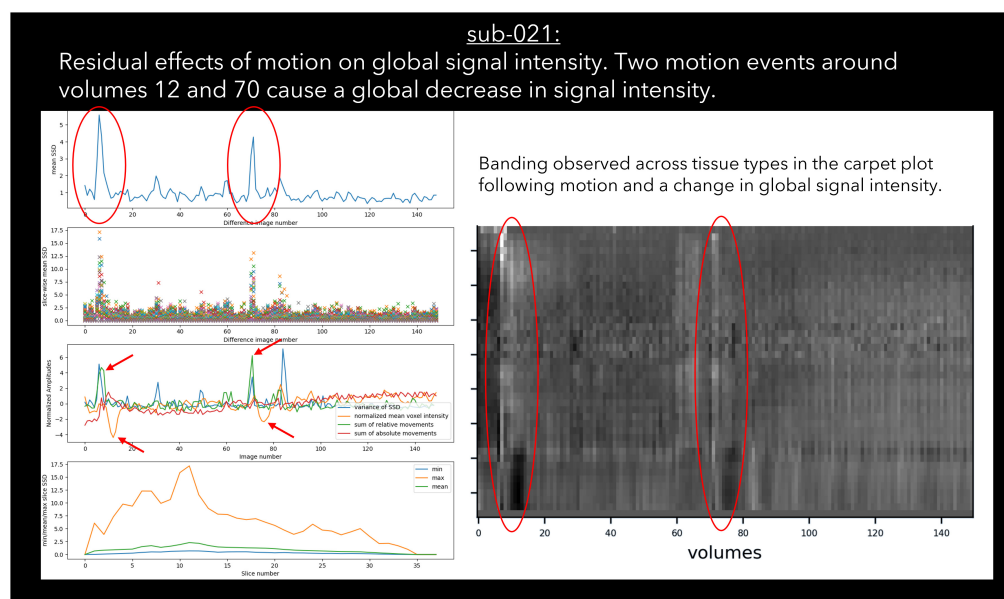


FIGURE 4

An example case of global signal loss following motion detected using output from pyfMRIqc.

preregistration of data quality control procedures and acceptable thresholds could also decrease the risk of biases influencing decision-making, while at the same time reducing questionable research practices more generally (Niso et al., 2022). However this has not yet been widely adopted in the neuroimaging community (Borghi and Gulick, 2018).

There are several limitations in the protocol and software as presented here. Firstly, our finding that often only a single rater classified a dataset as “uncertain” suggests that the quality control protocol presented (which is published unedited from its pre-assessment state), lacked nuance for interpreting edge cases that would otherwise have been classified as either “include” or “exclude.” Given that pyfMRIqc was initially designed to aid decision-making about the quality of raw/minimally pre-processed fMRI data, we suggest that future users err on the side of caution with respect to marking datasets for exclusion, and first fully pre-process data using their pipeline of choice and then determine whether this had a positive impact and reduced data quality issues. Second, cinnqc, and by extension pyfMRIqc, do not formally quantify the success of the minimal pre-processing steps. When designing software for users

with minimal programming experience, prioritising ease of use over functionality can reduce the freedom of more advanced users. For instance, brain extraction currently uses default arguments in FSL to identify brain and non-brain tissue (Smith, 2002). This process can sometimes exclude brain voxels (particularly at the boundary of the brain), or include non-brain voxels in the brain extracted image. However, these issues can be ameliorated *via* optional arguments that change the default values, but this requires fine tuning on a per-subject basis, or the use of other software like HD-BET or ANTsX (Isensee et al., 2019; Tustison et al., 2021). A method for integrating these features would improve the computational reproducibility of the quality control procedure, as currently users would need to generate these files separately and use the cinnqc nomenclature to integrate output with the rest of the pipeline. A third limitation is that pyfMRIqc does not currently provide visualisation for distributions of “no-reference” image quality metrics. As previously mentioned, MRIQC currently crowdsources these values from users by default to generate robust distributions (Esteban et al., 2017). Though pyfMRIqc does not currently have the userbase to make this an effective method for identifying outliers at the population level, visualising the distribution of these values for at least the group level would help users to make more informed decisions about the quality of data they have in their sample. Future versions of pyfMRIqc would be improved by focusing on including these features in the software, and could potentially integrate reference values from the MRIQC Web-API for equivalent metrics in a similar way to how MRIQCEPTION⁸ works.

In summary, we present a quality control protocol for pyfMRIqc (Williams and Lindner, 2020), implement it on data from the fMRI Open QC project (see text footnote 1), and assess its reliability using four independent raters. Data were classified by each rater as either “include,” “uncertain,” or “exclude,” based on the protocol and output



FIGURE 5

T1 weighted images for two cases of atypical brain structure that were present in the dataset.

⁸ <https://github.com/elizabethbeard/mriqception>

generated by pyfMRIqc and cinnqc, which automated minimal pre-processing, data curation, and identification of deviated acquisition parameters in the dataset. Our results indicate that our reliability between raters was good for “include” and “exclude” decisions, with κ values that ranged from moderate to substantial agreement. However, coefficients for the “uncertain” classification demonstrated little reliability between raters, and below chance level for all but one pair of raters. Furthermore, we found that in all but one cases where only one rater used the “uncertain” classification the other raters agreed with each other. We suggest that improvements in agreement between raters could be made by consulting sample-wide distributions of image quality metrics, increasing the clarity of the quality control protocol, and implementing further separate pre-processing steps before reassessing the data and deciding whether or not to exclude them.

Data availability statement

Each rater’s quality control report can be found in **Supplementary material as Data Sheets 3–6**, and at the University of Reading Research Data Archive <https://doi.org/10.17864/1947.000424>. pyfMRIqc and cinnqc output can be found on the cinnqc GitHub page <https://github.com/bwilliams96/cinnqc>.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

References

- Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034
- Arena, L., Morehouse, H. T., and Safir, J. (1995). MR imaging artifacts that simulate disease: How to recognize and eliminate them. *Radiographics* 15, 1373–1394. doi: 10.1148/radiographics.15.6.8577963
- Backhausen, L. L., Herting, M. M., Buse, J., Roessner, V., Smolka, M. N., and Vetter, N. C. (2016). Quality control of structural MRI images applied using freeSurfer—A hands-on workflow to rate motion artifacts. *Front. Neurosci.* 10:558. doi: 10.3389/fnins.2016.00558
- Benhajali, Y., Badhwar, A., Spiers, H., Urchs, S., Armoza, J., Ong, T., et al. (2020). A standardized protocol for efficient and reliable quality control of brain registration in functional MRI studies. *Front. Neuroinformatics* 14:7. doi: 10.3389/fninf.2020.00007
- Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Borghi, J. A., and Gulick, A. E. V. (2018). Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLoS One* 13:e0200562. doi: 10.1371/journal.pone.0200562
- Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., et al. (2018). The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* 32, 43–54. doi: 10.1016/j.dcn.2018.03.001
- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19:6. doi: 10.1038/mp.2013.78
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* 12:e0184661. doi: 10.1371/journal.pone.0184661
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76:378. doi: 10.1037/h0031619
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. J., and Turner, R. (1996). Movement-Related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312
- Gamer, M., Lemon, J., and Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement*. <https://CRAN.R-project.org/package=irr> (accessed October 14, 2022).
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:1. doi: 10.1038/sdata.2016.44
- Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* 1, 77–89. doi: 10.1080/19312450709336664
- Heiland, S. (2008). From A as in Aliasing to Z as in Zipper: Artifacts in MRI. *Clin. Neuroradiol.* 18, 25–36. doi: 10.1007/s00062-008-8003-y
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Author contributions

BW: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft, review and editing, visualization, and project administration. NH, CM, and GR: methodology, analysis, and review and editing. AC: resources, supervision, and review and editing. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2023.1070413/full#supplementary-material>

- Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., et al. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Hum. Brain Mapp.* 40, 4952–4964. doi: 10.1002/hbm.24750
- Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156. doi: 10.1016/S1361-8415(01)00036-6
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1016/S1053-8119(02)91132-8
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47:263. doi: 10.2307/1914185
- Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educ. Psychol. Meas.* 30, 61–70. doi: 10.1177/001316447003001015
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*, 2nd Edn. Thousand Oaks, CA: Sage.
- Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310
- Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., et al. (2013). Human connectome project informatics: Quality control, database services, and data visualization. *Neuroimage* 80, 202–219. doi: 10.1016/j.neuroimage.2013.05.077
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The openNeuro resource for sharing of neuroscience data. *Elife* 10:e71774. doi: 10.7554/eLife.71774
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochem. Med.* 22, 276–282. doi: 10.11613/BM.2012.031
- McNabb, C. B., Lindner, M., Shen, S., Burgess, L. G., Murayama, K., and Johnstone, T. (2020). Inter-slice leakage and intra-slice aliasing in simultaneous multi-slice echo-planar images. *Brain Struct. Funct.* 225, 1153–1158. doi: 10.1007/s00429-020-02053-2
- Niso, G., Botvinik-Nezer, R., Appelhoff, S., Vega, A. D. L., Esteban, O., Etzel, J. A., et al. (2022). Open and reproducible neuroimaging: From study inception to publication. *Neuroimage* 263:119623. doi: 10.1016/j.neuroimage.2022.119623
- Norton, M. I., Mochon, D., and Ariely, D. (2012). The IKEA effect: When labor leads to love. *J. Consum. Psychol.* 22, 453–460. doi: 10.1016/j.jcps.2011.08.002
- Polman, E. (2012). Self-other decision making and loss aversion. *Organ. Behav. Hum. Decis. Process.* 119, 141–150. doi: 10.1016/j.obhdp.2012.06.005
- Power, J. D. (2017). A simple but useful way to assess fMRI scan qualities. *Neuroimage* 154, 150–158. doi: 10.1016/j.neuroimage.2016.08.009
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Power, J. D., Plitt, M., Laumann, T. O., and Martin, A. (2017). Sources and implications of whole-brain fMRI signals in humans. *Neuroimage* 146, 609–625. doi: 10.1016/j.neuroimage.2016.09.038
- Pruim, R. H. R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., and Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage* 112, 267–277. doi: 10.1016/j.neuroimage.2015.02.064
- R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Raamana, P. R. (2018). *VisualQC: Assistive tools for easy and rigorous quality control of neuroimaging data*. Genè: Zenodo, doi: 10.5281/zenodo.1211365
- Rosen, A. F. G., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., et al. (2018). Quantitative assessment of structural image quality. *Neuroimage* 169, 407–418. doi: 10.1016/j.neuroimage.2017.12.059
- Smith, S. M. S. M. S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062
- Sokol-Hessner, P., Camerer, C. F., and Phelps, E. A. (2013). Emotion regulation reduces loss aversion and decreases amygdala responses to losses. *Soc. Cogn. Affect. Neurosci.* 8, 341–350. doi: 10.1093/scan/nss002
- Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., and Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5035–5040. doi: 10.1073/pnas.0806761106
- Tustison, N. J., Cook, P. A., Holbrook, A. J., Johnson, H. J., Muschelli, J., Devenyi, G. A., et al. (2021). The ANTsX ecosystem for quantitative biological and medical imaging. *Sci. Rep.* 11:1. doi: 10.1038/s41598-021-87564-6
- Van Dijk, K. R. A., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* 59, 431–438. doi: 10.1016/j.neuroimage.2011.07.044
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-minn human connectome project: An overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041
- Williams, B., and Lindner, M. (2020). pyfMRIqc: A software package for raw fMRI data quality assurance. *J. Open Res. Softw.* 8:1. doi: 10.5334/jors.280
- Zapf, A., Castell, S., Morawietz, L., and Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Med. Res. Methodol.* 16:93. doi: 10.1186/s12874-016-0200-9
- Zhang, Y., Brady, M., and Smith, S. M. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57.



OPEN ACCESS

EDITED BY

Tibor Auer,
University of Surrey, United Kingdom

REVIEWED BY

Thomas J. Ross,
National Institute on Drug Abuse (NIH),
United States
Joel Stoddard,
University of Colorado Anschutz Medical
Campus, United States

*CORRESPONDENCE

Joset A. Etzel
✉ jetzel@uwustl.edu

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroimaging

RECEIVED 14 October 2022

ACCEPTED 16 January 2023

PUBLISHED 17 February 2023

CITATION

Etzel JA (2023) Efficient evaluation of the Open
QC task fMRI dataset.

Front. Neuroimaging 2:1070274.

doi: 10.3389/fnimg.2023.1070274

COPYRIGHT

© 2023 Etzel. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Efficient evaluation of the Open QC task fMRI dataset

Joset A. Etzel*

Cognitive Control and Psychopathology Laboratory, Department of Psychological and Brain Sciences,
Washington University in St. Louis, Saint Louis, MO, United States

This article is an evaluation of the task dataset as part of the Demonstrating Quality Control (QC) Procedures in fMRI (FMRI Open QC Project) methodological research topic. The quality of both the task and fMRI aspects of the dataset are summarized in concise reports created with R, AFNI, and knitr. The reports and underlying tests are designed to highlight potential issues, are pdf files for easy archiving, and require relatively little experience to use and adapt. This article is accompanied by both the compiled reports and the source code and explanation necessary to use them.

KEYWORDS

fMRI, Quality Control, human, R, task

1. Introduction

This article is part of the Demonstrating Quality Control (QC) Procedures in fMRI (FMRI Open QC Project) methodological research project, and describes procedures for efficiently evaluating its task dataset. These procedures examine both the task (behavioral performance, stimuli presentation, etc.) and fMRI (motion, appearance of preprocessed images, etc.) aspects of the dataset. The code and criteria presented here are versions of that used for the Dual Mechanisms of Cognitive Control (DMCC; [Braver et al., 2021](#); [Etzel et al., 2022](#)) and multiple other projects in the Cognitive Control and Psychopathology Laboratory at Washington University in St. Louis (USA), and we hope will be useful and easily adapted by others.

QC procedures are often a balancing act between being so cursory that important problems are not identified, and so onerous that QC procedures are skipped entirely. The files and procedures presented here attempt to thread the needle; clearly highlighting the problems of greatest potential risk to the dataset and analysis integrity and validity, while remaining concise and easy to learn. These are intended to serve as a first step; a QC summary to allow efficient screening for potential issues, not to include all the details necessary to investigate any issues found.

These procedures are built around two dynamic report documents edited for the Open QC task fMRI dataset. The dynamic report framework is particularly well-suited to scientific programming because images, results, source code, and discussion are together in a single document. These reports are compiled to pdf files (convenient for archiving and have the same appearance wherever viewed), but there are many options for both output format and programming language. Regardless of the implementation details, I urge scientists to strive for clarity, simplicity, and stability when writing QC (or analysis) code over brevity and style purity, and hope that the documents included here can serve as a useful template.

2. Methods

One of the few statements a group of fMRI methods experts might all agree with is that there is a wide variety of methods for fMRI acquisition, processing, and analysis, none of which are unequivocally “best” for all (or even a specific) research questions. Given this methodological variety, quality judgments also widely diverge; the same images may be deemed suitable for one

analysis, but too flawed for another. There is also lack of consensus on which images to evaluate for quality, with some researchers using the raw images, others the preprocessed, and yet others a combination or after a processing pipeline used only for QC. Thus, one of the first decisions when approaching a new fMRI project is to determine which QC aspects are most relevant for the study, and how they will be evaluated.

In general, I believe the QC procedures should be dictated by each project's hypotheses and analyses, not by a standard protocol or fixed thresholds. Accordingly, I suggest performing QC on the images preprocessed as they will be for analysis. For example, if surface analyses are planned, the QC should include the surface reconstruction, and temporal mean, standard deviation (sd), and tSNR (temporal signal to noise ratio, here, mean/sd) images of the vertex timecourses (rather than the voxels used here). If a particular software package has been chosen, then the QC should be done using the images and motion parameters created by that package. Similarly, if images will be analyzed in subject space, the QC should be in subject space as well.

The reasoning behind this suggestion to perform QC on the preprocessed images is as a minimum, essential step; not to preclude other tests, but to maximize the likelihood of detecting an error arising anywhere in the pipeline. If a preprocessed image has high quality, its raw version is likely also of high quality, but a poor preprocessed image may or may not be the result of a low-quality raw image (e.g., if the participant moved during field mapping, warping may be introduced during the distortion correction preprocessing step). Again, I am not advising against including additional QC steps; procedures like evaluating image registration may be critically important in some cases. But I do advise that the preprocessed images always be examined for quality; that other image QC steps be in addition, not a replacement.

When considering task fMRI QC, participant behavior (e.g., task performance) is also of fundamental importance. Note that this is not evaluating whether the participant responded as theoretically predicted, but rather confirming that they were attempting to perform the instructed task (and not, say, sleeping or responding randomly). If the task requires frequent responses (e.g., button push and spoken word), response frequency may be useful as a proxy for attentiveness: long stretches without a response suggests the participant stopped performing the task. Other tasks may not require responses during the imaging session, but rely upon something like monitoring eye gaze or the results of a memory test performed after the session. Whatever the paradigm, for QC the aim is to determine a non-biased way to identify participants who did not have the minimally-valid task performance.

2.1. Data processing

The FMRI Open QC Project task dataset (Gorgolewski et al., 2017; Markiewicz et al., 2021) was provided in BIDS (Brain Imaging Data Structure; Gorgolewski et al., 2016) format specifically for QC demonstration, with the only guideline the assumption that the target analyses would be performed after spatial normalization to an MNI anatomical template and not include the cerebellum.

Given such minimal requirements, I chose to preprocess the images with fMRIPrep 21.0.1 (Esteban et al., 2019;

RRID:SCR_016216), which is reliable and straightforward to use, and has become our (and many other) group's default choice for fMRI preprocessing. Since surface analysis was not required, I chose to run fMRIPrep with volumetric preprocessing only, using the target MNI152NLin2009cAsym output template; the commands and generated text describing the preprocessing it performed are in the [Supplementary material](#). No other preprocessing was done before the image QC procedures described in this manuscript.

2.2. Resources

Two documents were prepared for QC assessment: one for the fMRI (openQC_fMRIQC), and the other for the stimuli and behavioral performance (openQC_behav). Both the compiled (.pdf) and source (.rnw) versions of each are available at <https://osf.io/ht543/>. These are dynamic report files, written in R (version 3.6.3, RRID:SCR_001905; R Development Core Team, 2020) and knitr (version 1.39; Xie, 2014, 2015, 2022); all code is contained within the source (.rnw) versions of each file. The documents depend upon the RNIfti (version 1.3.0; Clayden et al., 2020) and fields (version 11.6; Nychka et al., 2017) R packages, as well as AFNI 22.0.11 (RRID:SCR_005927; Cox, 1996; Cox and Hyde, 1997).

The task timing and responses in openQC_behav were read directly from the provided _events.tsv files. Similarly, openQC_fMRIQC read the six motion regressors and framewise displacement (FD) directly from the _desc-confounds_timeseries.tsv files produced by fMRIPrep (columns trans_x, trans_y, trans_z, rot_x, rot_y, rot_z, and framewise_displacement). The voxelwise temporal mean, standard deviation (sd), and tSNR (mean/sd) images were calculated with AFNI 3dTstat and 3dcalc functions, using the entire run (without censoring); see the startup code chunk in openQC_fMRIQC.rnw. While the number of censored frames (with FD above threshold) is included in the QC criteria as detailed below, I prefer not to censor when calculating the temporal mean, sd, and tSNR images during QC, to visually exaggerate differences between runs.

2.3. QC criteria

Four criteria necessary for task fMRI QC are presented below and applied to the FMRI Open QC Project task dataset. I want to emphasize that these are (in my opinion) necessary criteria, but not sufficient for all cases, nor a comprehensive list of all aspects of dataset quality. Indeed, while preparing this manuscript a reviewer commented that no criteria involved checking the raw (before preprocessing) anatomical or functional images. In our ongoing projects such a criterion is actually used: the experimenter rates the quality of the anatomical images immediately after acquisition, so that poor-quality scans can be repeated (<https://osf.io/a7w39/>). We no longer routinely evaluate the raw functional images, since when we have performed such checks they seem to add time and complexity without identifying issues beyond than those also found with the preprocessed images (Criterion D below). This decision to focus QC on preprocessed images is a judgement made for our particular

research aims and resource limitations; please consider what is most important in your situation.

2.3.1. Criterion A: Excessive motion

It can be surprisingly difficult to quantify how much motion is “excessive,” especially in fMRI datasets with high apparent motion (Inglis, 2016; Etzel, 2017a; Power et al., 2019; Fair et al., 2020). For task fMRI we have adapted the procedure described in Siegel et al. (2014), and censor individual frames with FD > 0.9. Further, if more than 20% of the frames in a run are censored, then the entire run is omitted (Etzel, 2017b; Etzel et al., 2022). While these are quantitative thresholds, I advise also viewing plots of the motion regressors during QC (first part of openQC_fmriQC.pdf), and not solely rely on a count of censored frames or other summary statistic, since respiratory task entrainment, forceful blinking, machine vibrations, and many other things can cause unexpected (and potentially problematic) patterns in the motion regressors. We do not generally exclude runs or participants for an unusual motion pattern alone, but such patterns should be monitored as part of routine QC, since they may indicate that a problem is developing with image acquisition (e.g., a hardware fault), or help inform analysis strategy (e.g., if have respiratory task entrainment, including many motion regressors in the GLMs may remove substantial task information).

The censoring threshold of FD > 0.9 suggested here is much more lenient than advised by many researchers (including Siegel et al., 2014, which suggests 0.5 for typical adults), though we have found it a useful starting point. The choice of censoring threshold and method (e.g., on FD, enorm, or translation; single frame or adjacent as well) depends on multiple factors, perhaps most importantly study design and planned analysis. If temporal correlations will be used (e.g., for functional connectivity analyses), stringent motion thresholds and filtering techniques are essential (Satterthwaite et al., 2013; Ciric et al., 2017, 2018). With task designs, higher motion levels may be tolerable, if not strongly linked to trial types. The linkage of (apparent) motion and trial timing is common (Perl et al., 2019), however, and poses a serious methodological challenge. Plotting trial onsets with the motion regressors (as in openQC_fmriQC.pdf) can aid in spotting potentially significant confounding of task and motion, but much work remains to be done in this area.

2.3.2. Criterion B: Improper task presentation

To estimate task-related responses consistently across participants we generally need approximately the same amount of imaging data from each participant, so this criterion is to exclude a participant if <½ of their trials have usable data (in the sense of being analyzable). Given the wide variety of task paradigms, the definition of “usable” data also varies, but at minimum both the fMRI images and task presentation details (e.g., stimulus onset time) must be present for the trials to be usable. For examples of the types of cases that may lead to this criterion being met, consider that incomplete task runs may result from hardware failure (e.g., projector bulb breaks; the participant mentions after scanning that they did not hear the audio stimuli), participant request (e.g., they ask to end a run early), or presentation error (e.g., the experimenter started the wrong task script; the task was programmed incorrectly and did not present the necessary trials).

While not implemented here, fMRI images for the run being present is not sufficient for a particular trial to be analyzable: it may have occurred during a period of excessive motion, and so be censored (which removes the affected frames from analysis). There is accordingly an interaction between criteria A and B: if a participant has many frames above the censoring threshold, the timing between the task trials and censored frames should be evaluated, as not all frames have the same impact. For example, some participants tend to move outside of task blocks (e.g., at the end of a run), which will change the number of analyzable trials less than if the motion occurs during the trials themselves.

2.3.3. Criterion C: Invalid task performance

Note that this is not excluding participants who performed the task “incorrectly” according to the experimental hypotheses, but rather those who did not perform the expected task at all, such as not following the instructions or attending to the stimuli. For example, if the task involves responding to visual stimuli, we want to exclude participants who fell asleep or closed their eyes throughout stimulus presentation. Given the wide variety of protocols and priorities there is no universally applicable way to describe valid task performance; the most important aspects of each experiment should be considered, and criteria incorporate features like catch trials or eye gaze if present.

In the FMRI Open QC Project task dataset we only have the task information that can be gleaned from the BIDS events.tsv files; far less than is typical. Proceeding nevertheless, it appears that participants were asked to make a button-press response after every trial, the trials were fairly short and rapid (seven or more each minute; openQC_behav.pdf), and most participants responded accurately to most trials. In these types of designs it can work well to define invalid task performance quantitatively by no-response trials: exclude if a participant fails to respond to five or more trials in a row or more than 40 percent of the total trials within a run [thresholds adopted from the HCP task protocols (WU-Minn Consortium of the NIH Human Connectome Project, 2013)]. Note that this criterion is not of correct trial responses, but of any trial response (in cases where every trial requires a response).

Qualitatively, the responses should be reviewed for unambiguous patterns indicating that the participant was not performing the task correctly, such as using only one response button or responding in a repetitive sequence instead of to the stimuli.

The motivation for including quantitative thresholds is the need to distinguish inattention from poor task performance in the most unbiased way possible. Assuming the experimenters wish to include participants with a range of performance, people finding the task difficult will generally have a mix of correct- and incorrect-response trials, and slower RTs than people finding it easy. If the task requires a response to be made within a certain amount of time, slower RTs can lead to some trials not have a recorded response, even though the person was attentive and trying to perform the task. Thus, we may interpret 10 no-response trials differently if they were scattered evenly throughout the run (suggesting task difficulty, particularly if accompanied by low accuracy or slow RTs) than in a group of sequential trials (suggesting inattentiveness, particularly if trials with responses tend to be fast and accurate).

Plots such as in Figure 3 and careful examination of response patterns in pilot data or previous experiments may assist in setting the quantitative thresholds for a particular experiment. The

TABLE 1 Task-based fMRI QC criteria: exclude the run for a subject if:

	Name	Type	Details
A	Excessive motion	Quantitative	20% or more of the frames have more than 0.9 mm FD
B	Improper task presentation	Quantitative	Less than half of the trials have usable data (e.g., due to hardware failure).
C	Invalid task performance (e.g., participant fell asleep)	Quantitative and qualitative	There is no response for five or more trials in a row or more than 40% of the total. Also exclude if the pattern of responses indicates unambiguously invalid performance (e.g., only one response button used).
D	Failed image acquisition and/or preprocessing	Qualitative	The preprocessed temporal mean, sd, and tSNR images do not resemble the MNI anatomical template (e.g., distorted shape), have the expected properties (e.g., the sd image does not resemble an arteriogram), have unusual structured noise, or are otherwise clearly and excessively affected by artifacts.

appropriateness of the quantitative thresholds of five or more no-response trials in a row or 40% of the total can't be evaluated in this case (given the lack of experimental details), but can serve as a default or starting point. While any threshold is imperfect, this may pose a smaller risk than that of experimenter bias if only qualitative criteria are used to determine which participants to exclude.

2.3.4. Criterion D: Failed image acquisition and/or preprocessing

For this criterion, qualitatively review temporal mean, sd, and tSNR (mean/sd) images of each run, looking for incorrect or unusual cases requiring further investigation. Visual arrays with multiple runs side-by-side assist in evaluating typical variability, and thus also in spotting exceptions. We have found it useful to concentrate the initial QC evaluation on a few easy-to-spot features. First, check the volumetric temporal mean images for “alien” or “escaping” brains. No preprocessed image will exactly match the anatomic template, but distortions should not be extreme (“aliens”), and the brain should always be centered in the same part of the image (not “escaping” the frame). Second, the mean volumetric images should have clearly visible brain structure (i.e., resemble an anatomical scan), while the sd images should be brightest around the edges and in large vessels. Throughout, the images should be examined for unusually structured noise, dark areas, or other oddities. Surface images are more difficult to qualitatively review, since they are typically plotted on a single surface underlay and only include the gray matter ribbon. However, a useful QC feature is to look for the central sulcus in the temporal mean images, which should be clearly visible as “tiger stripes” at the top of each hemisphere; non-anatomical dark patches or “polka dot” patterns should also be noted.

If something is spotted during these qualitative checks of the statistical images, the run should be investigated in detail before deciding whether or not to exclude it. For example, if the raw (unprocessed) images appear as expected but the preprocessed images do not, an error likely occurred during preprocessing and may be possible to correct. If the raw images are also affected, then the run is likely unusable, and the source of the problem should be investigated to see if its recurrence can be avoided. Sometimes it is unclear whether an unusual run should be included or not, such as when dropout or noise is only slightly higher than typical. In these uncertain cases it can be useful to evaluate whether the results of positive control analyses (e.g., of strong effects such as button presses; Niso et al., 2022) are within the range of other participants, and exclude if not.

3. Results

Applying these criteria to the Open QC dataset, in my judgment three participants should be excluded from the hypothetical analysis: one for failed image acquisition (Table 1 criterion D and sub-010), and two for invalid task performance (Table 1 criterion C, sub-016, and sub-025). The others vary in quality, particularly of the images, but do not reach the exclusion thresholds. openQC_fMRIQC.pdf and openQC_behav.pdf (<https://osf.io/ht543/>) contain the necessary figures and statistics to evaluate the Open QC task dataset in terms of the Table 1 criteria, as will be described here.

3.1. Results from applying the image-related criteria

The first document, openQC_fMRIQC.pdf, is intended to highlight key image-related features of the task fMRI runs: motion (criterion A) and success of acquisition and preprocessing (criterion D). The first section has line plots of the six realignment parameters and FD for each run, with trial onsets, censoring threshold (FD > 0.9), and censored frames marked. The number and proportion of censored frames are printed on each plot, for ease of comparing to the 0.2 censoring exclusion criterion. No participants reached this threshold, and while movement clearly varies across participants, I do not suggest excluding any due to excessive (or highly unusual) movement. Interestingly, the degree of apparent motion varies across participants; for example minimal in sub-011 and sub-030, but clear in sub-012 and sub-022. Some participants have brief instances of overt head motion, such as sub-003 and sub-018. Overall, the FD > 0.9 frame censoring threshold seems reasonable for this dataset, appropriately identifying the larger overt, but not apparent, head motion.

The second section of openQC_fMRIQC.pdf has plots of the temporal mean, standard deviation, and tSNR (mean/sd) images for each participant, for applying criterion D (or more concretely, looking for oddities; images that are not like the others). As shown in Figure 1, while all other participants' images resemble the MNI preprocessing target template, sub-010 is clearly a different shape. Other than the distorted sub-010 images, the summary images have the expected characteristics (e.g., anatomical structures are visible on the means; sd have bright vessels). To further evaluate sub-010, I looked at the provided raw image file (sub-010_task-pamenc_bold.nii.gz); if the raw image looks typical, we could suspect

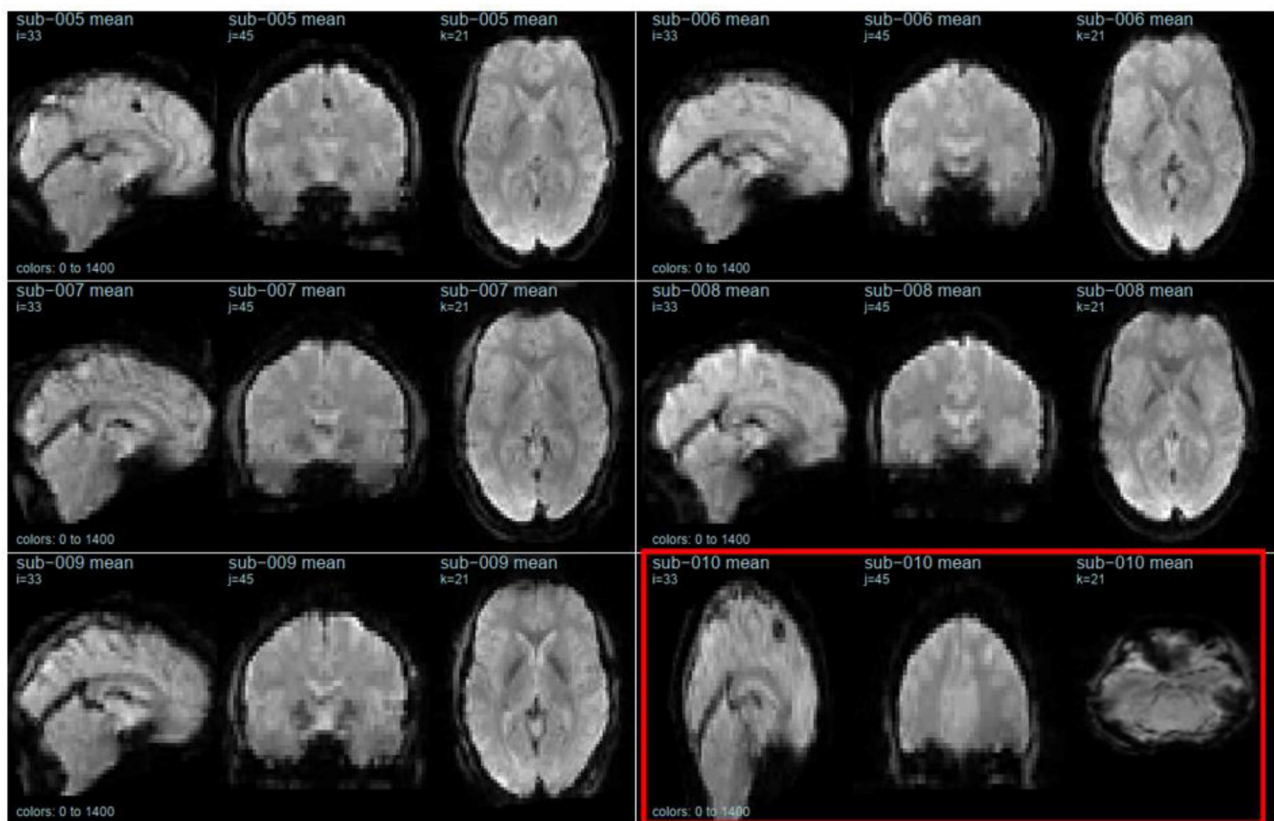


FIGURE 1
Temporal mean images for six participants, calculated after preprocessing. sub-010 (outlined in red) is highly distorted. This figure is from page 12 of openQC_fmriQC.pdf.

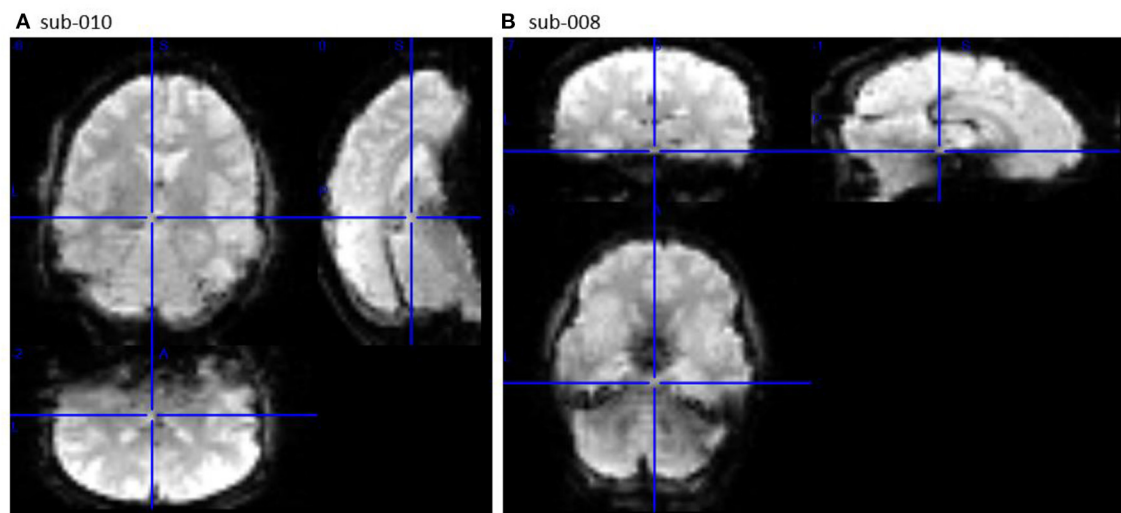


FIGURE 2
Frame 100 of the raw image timeseries (`_bold.nii.gz`) for sub-010 [left, (A)] and comparison sub-008 [right, (B)], viewed in MRIcron (Rorden et al., 2007; RRID:SCR_002403). sub-010 was not acquired with the same parameters as sub-008 (and the other participants).

that the problem was introduced during preprocessing. However, here, the problem is present in the raw image as well: **Figure 2** shows frame 100 from sub-010 on the left, and for a comparison example, sub-008 on the right. The image orientation and planes are clearly

different for sub-010, so the unusual appearance in **Figure 1** was not introduced by preprocessing. Further, the phase encoding direction and other fields in `sub-010_task-pamenc_bold.json` vary from the other participants. We can thus conclude that the image acquisition

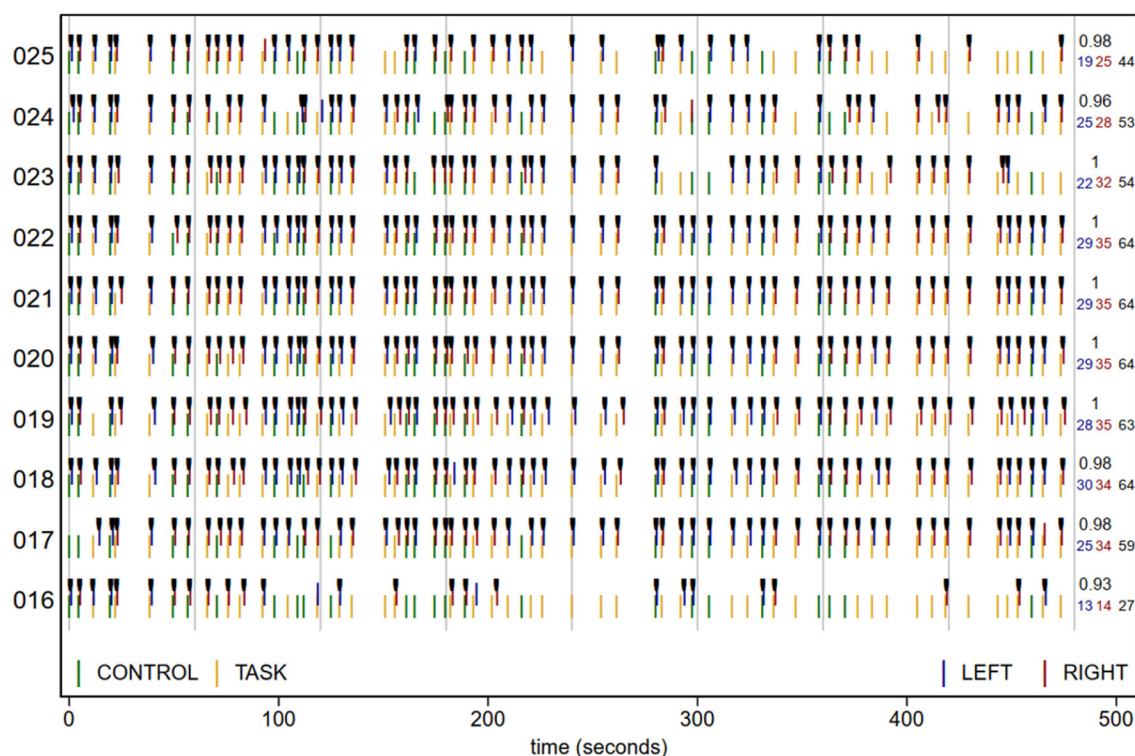


FIGURE 3

Task trial onset time (colored by trial type), response (at onset + reaction time), and accuracy (black ticks) for 10 participants. Vertical gray lines are at 1-min intervals. Numbers at right margin give the number of LEFT (blue) and RIGHT (red) responses, total responses (black; 64 if no trials lacked a response), and proportion correct of the trials with a response. sub-016 responded correctly to all trials in the 1st min, but then had more and more trials without a response, suggesting that they became less attentive as the run progressed. sub-025 was also excluded due to criterion C, and while their strings of no-response trials did not reach the 5-trial threshold until the last minute of the run, they missed noticeably more trials in the second than first half of the run.

was incorrect for sub-010, and the participant's imaging data should be excluded.

In some cases the raw images first appear odd, but are recoverable (e.g., by changing parameters or preprocessing template). Other issues arise from errors that causes the images to have fundamentally different properties than the rest of the dataset (e.g., if the wrong head coil or encoding direction was used), and so must be excluded. If this was an ongoing experiment, the researchers should investigate how it came about that the wrong acquisition protocol was used for the session, and if changes to the SOPs [Standard Operating Procedures (Etzel et al., 2022; Niso et al., 2022), <https://osf.io/6r9f8>] could avoid the mistake happening again.

3.2. Results from applying the task-related criteria

The second document, openQC_behav.pdf, is intended to highlight and evaluate key aspects of the task presentation (criterion B) and behavioral performance (criterion C). The code in chunk code2 counts how many trials of each type were presented to each participant, and prints an error message if the counts are not as expected. For this dataset, it checks the number of CONTROL and TASK trials in each run, and since all participants have the same number, no participants were excluded for criterion B. If some aspect

of the task paradigm is key for valid analysis (e.g., each stimulus must be presented exactly twice), this should be explicitly tested in this section, and any violations clearly highlighted.

The plot in openQC_behav.pdf, excerpted in Figure 3, summarizes the task presentation and performance for each participant (y axis). Time is along the x axis, and each green (CONTROL) and yellow (TASK) plotting symbol indicates the type and onset time of a trial (read from the origcopy _events.tsv files). The blue (LEFT) and red (RIGHT) lines show the time and type of each button press, with black tick marks on correct responses. The numbers in the right margin list the number of LEFT and RIGHT responses, their total, and the proportion correct of trials with a response. While dense, with practice a lot of task and performance information can be quickly scanned in plots like these, including trial timing and randomization (e.g., here we can see that all participants had the same trial order and timing), and unexpected response patterns.

To reduce the chances of missing an exception, the quantitative task performance criteria (C) are tested explicitly in chunk code2. Three notifications are printed: that sub-016 has both >40% no-response trials and 5 or more no-response trials in a row, and that sub-025 has 5 or more no-response trials in a row. These strings of trials without a response are visible in the participants' rows in Figure 3, as stretches of trial onsets (green and yellow lines) without the corresponding responses (red and blue lines). Accordingly, both sub-016 and sub-025 should be excluded from

analysis due to excessive missing responses. We can also observe that participants made very few errors in this experiment; nearly all responses that were made, were correct (sub-013 is lowest at 0.88 accuracy). In some paradigms or analyses it may be relevant to establish additional criteria, such as excluding participants with accuracy below a threshold.

4. Discussion

This article presented an evaluation of the fMRI Open QC Project task dataset, as part of the Demonstrating Quality Control Procedures in fMRI methodological research project. Both the task and fMRI aspects of the dataset were examined, applying the criteria summarized in Table 1 via the figures and statistics in the two dynamic report summary documents (openQC_fmriQC.pdf and openQC_behav.pdf; R, AFNI, and LaTeX) available at <https://osf.io/ht543/>. Using these criteria, I suggest that three participants should be excluded from the hypothetical analysis: one for failed image acquisition and two for invalid task performance.

I do not believe that there is a “perfect” or even “ideal” procedure for QC in psychological or neuroimaging research: new potential issues are identified continually, and the sheer amount of data makes checking every piece impossible. Nevertheless, there clearly is a terrible way to do QC: by omission. We have likely all been involved in a project where a critical artifact or error was discovered late, sometimes so severe that the dataset must be abandoned or a publication corrected.

Since the fMRI Open QC Project task dataset was complete (acquisition finished years ago) and small (only one run per person), I included all of the participants in each of the two QC summary documents. This is only appropriate on completed datasets, however. For new and ongoing projects, QC summary documents should be created for each participant on a continual basis, and reviewed as soon after each session as possible, a task made efficient by dynamic reports and clear guidance on how to review the reports [examples of such single-subject QC reports from the DMCC project (Braver et al., 2021; Etzel et al., 2022) are at <https://osf.io/7xkq9> and <https://osf.io/z62s5>]. While no one can guarantee that such ongoing QC procedures will prevent disaster, they can certainly help reduce the odds of collecting an unusable dataset, by allowing researchers to catch serious issues early, when they can still be corrected.

The material presented here is intended to serve both as inspiration and a template for adapting to your own datasets. The code in the summary documents is designed to be straightforward and approachable; easy to edit for other datasets or reimplement in a new language. A number of QC software packages which can generate reports without programming are also available, including MRIQC (Esteban et al., 2017). However, accomplished, I suggest QC include reviewing the images themselves, not only summary statistics.

Particularly with task fMRI, but to some extent with any study of living participants, both the task/behavior and imaging parts of the dataset need to be included in the QC. Which aspects and criteria are most important will vary with each

dataset and analysis, so I suggest starting by considering which features absolutely must be true for the analyses and inferences to be valid, and ensuring that those features are covered in the QC procedures. We may not be able to achieve “perfect” QC, analysis, or results, but good QC procedures can let us be confident that what we are analyzing and reporting is real, not wholly invalid.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: <https://osf.io/qaesm/>.

Author contributions

JE wrote the manuscript and code and conducted the analyses.

Funding

This work was supported by the National Institutes of Health, grant number: R37MH066078 to Todd Braver.

Acknowledgments

I thank Rebecca Feldman and Maya Quale for assistance in running fMRIPrep and commenting on the manuscript.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnimg.2023.1070274/full#supplementary-material>

References

- Braver, T. S., Kizhner, A., Tang, R., Freund, M. C., and Etzel, J. A. (2021). The dual mechanisms of cognitive control project. *J. Cogn. Neurosci.* 2021, 1–26. doi: 10.1162/jocn_a_01768
- Ciric, R., Rosen, A. F. G., Erus, G., Cieslak, M., Adebimpe, A., Cook, P. A., et al. (2018). Mitigating head motion artifact in functional connectivity MRI. *Nat. Protocols* 13, 2801–2826. doi: 10.1038/s41596-018-0065-y
- Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage* 154, 174–187. doi: 10.1016/j.neuroimage.2017.03.020
- Clayden, J., Cox, R. W., and Jenkinson, M. (2020). *RNifti: Fast R and C++ Access to NIfTI Images*. Available online at: <https://CRAN.R-project.org/package=RNifti> (accessed January 23, 2023).
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Cox, R. W., and Hyde, J. S. (1997). Software tools for analysis and visualization of FMRI data. *NMR Biomed.* 10, 171–178. doi: 10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from Unseen Sites. Edited by Boris C Bernhardt. *PLoS ONE* 12, e0184661. doi: 10.1371/journal.pone.0184661
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). FMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* 16, 111–116. doi: 10.1038/s41592-018-0235-4
- Etzel, J. A. (2017a). *Task FMRI Motion Censoring (Scrubbing) #1: Categorizing MVPA Meanderings (blog)*. Available online at: <http://mvpa.blogspot.com/2017/04/task-fmri-motion-censoring-scrubbing-1.html> (accessed April 21, 2017).
- Etzel, J. A. (2017b). *Task FMRI Motion Censoring (Scrubbing) #2: Implementing MVPA Meanderings (blog)*. Available online at: <https://mvpa.blogspot.com/2017/05/task-fmri-motion-censoring-scrubbing-2.html> (accessed May 4, 2017).
- Etzel, J. A., Brough, R. E., Freund, M. C., Kizhner, A., Lin, Y., Singh, M. F., et al. (2022). The dual mechanisms of cognitive control dataset, a theoretically-guided within-subject task FMRI battery. *Sci. Data* 9, 114. doi: 10.1038/s41597-022-01226-4
- Fair, D. A., Miranda-Dominguez, O., Snyder, A. Z., Perrone, A., Earl, E. A., Van, A. N., et al. (2020). Correction of respiratory artifacts in MRI head motion estimates. *NeuroImage* 208, 116400. doi: 10.1016/j.neuroimage.2019.116400
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3, 160044. doi: 10.1038/sdata.2016.44
- Gorgolewski, K. J., Durnez, J., and Poldrack, R. A. (2017). Preprocessed consortium for neuropsychiatric phenomics dataset. *F1000Research* 6, 1262. doi: 10.12688/f1000research.11964.2
- Inglis, B. (2016). *Respiratory Oscillations in EPI and SMS-EPI. PractiCal FMRI: The Nuts & Bolts (blog)*. Available online at: <https://practicalfmri.blogspot.com/2016/10/respiratory-oscillations-in-epi-and-sms.html> (accessed October 7, 2016).
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *ELife* 10, e71774. doi: 10.7554/eLife.71774
- Niso, G., Botvinik-Nezer, R., Appelhoff, S., Vega, A. D. L., Esteban, O., Etzel, J. A., et al. (2022). Open and reproducible neuroimaging: From study inception to publication. *NeuroImage* 2022, 119623. doi: 10.1016/j.neuroimage.2022.119623
- Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). *Fields: Tools for Spatial Data*. Boulder, CO: University Corporation for Atmospheric Research.
- Perl, O., Ravia, A., Rubinson, M., Eisen, A., Soroka, T., Mor, N., et al. (2019). Human non-olfactory cognition phase-locked with inhalation. *Nat. Hum. Behav.* 3, 501–512. doi: 10.1038/s41562-019-0556-z
- Power, J. D., Lynch, C. J., Silver, B. M., Dubin, M. J., Martin, A., and Jones, R. M. (2019). Distinctions among real and apparent respiratory motions in human FMRI data. *NeuroImage* 201, 116041. doi: 10.1016/j.neuroimage.2019.116041
- R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: <http://www.r-project.org> (accessed January 23, 2023).
- Rorden, C., Karnath, H.-O., and Bonilha, L. (2007). Improving lesion-symptom mapping. *J. Cogn. Neurosci.* 19, 1081–1088. doi: 10.1162/jocn.2007.19.7.1081
- Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage* 64, 240–256. doi: 10.1016/j.neuroimage.2012.08.052
- Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., et al. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points: Censoring high motion data in FMRI. *Hum. Brain Map.* 35, 1981–1996. doi: 10.1002/hbm.22307
- WU-Minn Consortium of the NIH Human Connectome Project (2013). *WU-Minn HCP Q2 Data Release: Reference Manual Appendix IV – HCP Protocol Standard Operating Procedures*. WU-Minn Consortium of the NIH Human Connectome Project. Available online at: https://www.humanconnectome.org/storage/app/media/documentation/data_release/Q2_Release_Appendix_IV.pdf (accessed January 23, 2023).
- Xie, Y. (2014). “Knitr: A comprehensive tool for reproducible research in R,” in *Implementing Reproducible Research, The R Series*, eds V. Stodden, L. Friedrich, and R. D. Peng (Boca Raton, FL: CRC Press, Taylor & Francis Group). 5–28.
- Xie, Y. (2015). *Dynamic Documents with R and Knitr. 2nd Edn*. Boca Raton, FL: CRC Press/Taylor & Francis.
- Xie, Y. (2022). *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. Available online at: <https://yihui.name/knitr/> (accessed January 23, 2023).



OPEN ACCESS

EDITED BY

Paul A. Taylor,
National Institute of Mental Health (NIH),
United States

REVIEWED BY

Stefano Moia,
Center for Biomedical Imaging (CIBM),
Switzerland
Hua Xie,
University of Maryland, College Park,
United States

*CORRESPONDENCE

Bin Lu
✉ lub@psych.ac.cn
Chao-Gan Yan
✉ yancg@psych.ac.cn

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

RECEIVED 14 October 2022

ACCEPTED 01 February 2023

PUBLISHED 21 February 2023

CITATION

Lu B and Yan C-G (2023) Demonstrating
quality control procedures for fMRI in DPABI.
Front. Neurosci. 17:1069639.
doi: 10.3389/fnins.2023.1069639

COPYRIGHT

© 2023 Lu and Yan. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Demonstrating quality control procedures for fMRI in DPABI

Bin Lu^{1,2*} and Chao-Gan Yan^{1,2,3,4*}

¹CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China, ²Department of Psychology, University of Chinese Academy of Sciences, Beijing, China, ³International Big-Data Center for Depression Research, Institute of Psychology, Chinese Academy of Sciences, Beijing, China, ⁴Magnetic Resonance Imaging Research Center, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

Quality control (QC) is an important stage for functional magnetic resonance imaging (fMRI) studies. The methods for fMRI QC vary in different fMRI preprocessing pipelines. The inflating sample size and number of scanning sites for fMRI studies further add to the difficulty and working load of the QC procedure. Therefore, as a constituent part of the Demonstrating Quality Control Procedures in fMRI research topic in Frontiers, we preprocessed a well-organized open-available dataset using DPABI pipelines to illustrate the QC procedure in DPABI. Six categories of DPABI-derived reports were used to eliminate images without adequate quality. After the QC procedure, twelve participants (8.6%) were categorized as excluded and eight participants (5.8%) were categorized as uncertain. More automatic QC tools were needed in the big-data era while visually inspecting images was still indispensable now.

KEYWORDS

quality control, fMRI, neuroimaging, DPABI, pipeline

1. Introduction

Quality control (QC) is an important stage for functional magnetic resonance imaging (fMRI) studies. Images with a variety of artifacts, noticeable head motion artifacts, a low signal-to-noise ratio, inadequate slices, etc., are eliminated by researchers. Some nuisance signals such as head motion artifacts would be further regressed out and included as covariates in the following statistic. In the present study, we illustrated the fMRI quality control routine in DPABI by preprocessing a well-organized fMRI dataset.

Quality control for fMRI is becoming more challenging at this point. The challenge stems from several sources. First, to reduce the false positive rate and increase the reproducibility of an fMRI experiment, the sample size required has significantly improved over the past decade. More MRI data result in increased human power consumption in the non-automatic QC procedures such as visually screening the T1-weighted images with unacceptable motion artifacts (Backhausen et al., 2016). Second, even if the workload of researchers has been lessened by well-known preprocessing tools like fMRIPrep (Esteban et al., 2019), C-PAC (Michael et al., 2013), and DPABI (Yan et al., 2016), the optimum quality control procedures in these preprocessing pipelines still call for human involvement in the process. Several

fully automatic brain MRI QC tools have been developed but the generalizability of them needs to be further validated on the independent datasets (Mortamet et al., 2009; Alfaro-Almagro et al., 2018; Bastiani et al., 2019). Third, the generalizability of findings drawn from multi-center image acquisition studies could be significantly improved. However, the variability across MR manufacturers, scanning procedures, daily scanner QC standards, and other factors may prevent researchers from applying a consistent criterion to exclude data. Therefore, a meta-data report for all the preprocessed participants would contribute to avoiding mistakes such as deficiency of time points in functional sessions or abnormal TR. In general, the present QC tools are designed to reduce the mechanically repetitive operations of users by providing and illustrating more user-friendly quality assessments. These tools may significantly alleviate the working load added by increased sample size and multi-center design, but could not replace the decision-making procedure of human beings in QC. Last but not least, the open-science data-sharing trend offers an unpretentious opportunity to reuse existing data or combine a vast number of images to carry out ambitious large-scale analyses. However, the inclusion of meta-data of samples could be various among different datasets and acquisition parameters might be unavailable for some datasets. Even worse, some flaws can be hard for users of these open datasets to identify (e.g., the flipped left-right direction, redundant images for an MR series, wrong participant sex labels, etc.). To summarize, the issues raised above demand that researchers prioritize the quality control procedure and integrate more efficient and user-friendly tools into preprocessing pipelines.

Most of the popular fMRI pipelines have their unique QC routines. The MRIQC is a pioneer specialized QC framework that incorporates a variety of techniques (Esteban et al., 2017). In recent, the main contributors of MRIQC developed another important pipeline fMRIPrep for fMRI preprocessing. The fMRIPrep would produce a series of intuitive dynamic graphs and charts to demonstrate the effectiveness of Bold-T1 image co-registration, brain surface reconstruction, spatial normalization, and the severity of head motion after fMRI preprocessing. These graphs and reports are frequently invoked by QC procedures in the other pipelines such as DPABISurf (Yan et al., 2021) and ENIGMA HALFPipe (Waller et al., 2022). For example, HALFPipe provides an interactive webpage for users to evaluate an integrated quality report derived from fMRIPrep and other tools for each participant. And DPABI also combines all the reports from a group of participants into three reports to reduce repetitive operations. As mentioned above, QC was essential for large-scale, multi-center imaging projects. Therefore, the recent large-scale projects like UKBiobank (Alfaro-Almagro et al., 2018), ABCD (Hagler et al., 2019), and ENIGMA (Waller et al., 2022) also created their own (combination of) QC methods. In addition to these specialized QC tools, imaging formatter such as DCM2NIIX (Li et al., 2016), BIDS-validator and DPABI_InputPreparer could also be used to check for the absence of imaging meta-data in QC. DPABI is a widely-used user-friendly toolbox for fMRI data processing. Both existing QC tools and in-house QC procedures have been integrated into the volume-based pipeline DPARSF, surface-based pipeline DPABISurf and specialized QC modules. The purpose of this work was to demonstrate how to QC fMRI data in DPABI. Participants with poor image quality

were excluded based on a set of criteria which was thoroughly described.

2. Materials and methods

2.1. Participants

A collection of resting-state fMRI data, called fmri-open-qc-rest, was used for demonstrating the QC procedure in DPABI. The fmri-open-qc-rest dataset includes participants pooled from 7 different datasets, each with about 20 subjects (total $N = 139$). It's a demonstrating data of the fMRI Open QC Project and the anonymous samples were selected from widely-used open-available datasets such as the functional connectome project (FCP) (Biswal et al., 2010), the autism brain imaging data exchange (ABIDE) (Di Martino et al., 2014) and the OpenNeuro resource (Markiewicz et al., 2021). The sex and age of participants were not available in the fmri-open-qc-rest dataset.

2.2. Surface-based MRI preprocessing

Both a volume-based pipeline (DPARSF) and a surface-based pipeline (DPABISurf) in DPABI were used to preprocess the MRI data. Surface-based methods are increasingly common in the most recent studies and are superior to volume-based methods in terms of structure localization, spatial smoothing, and reproducibility (Coalson et al., 2018). However, the surface-based methods were time-consuming and omitted the analysis of subcortical and cerebellar areas. The volume-based approaches would be appropriate for conducting whole-brain analysis, preprocessing large datasets, etc. Additionally, the DPARSF pipelines reorient/QC module offered a user-friendly graphical user interface for visually assessing the image quality before the remaining laborious stages (e.g., structure segmentation).

In specific, surface-based preprocessing was performed by DPABISurf (Yan et al., 2021), a surface-based fMRI data analysis toolbox evolved from DPABI/DPARSF. DPABISurf used docker technology to wrap the whole computing environment for fMRIPrep (Esteban et al., 2019), FreeSurfer (Fischl, 2012), ANTs (Tustison et al., 2014), FSL (Jenkinson et al., 2012), AFNI (Cox, 1996), SPM (Ashburner, 2012), GNU Parallel (Tange, 2011), PALM (Winkler et al., 2014), MATLAB (The MathWorks Inc., Natick, MA, USA), Docker¹ and DPABI (Yan et al., 2016), etc. The pipelines mentioned above have their own preprocessing and QC procedures and an elaborate comparison among these pipelines could be found in the ENIGMA HALFPipe references (Waller et al., 2022). The resting-state functional images and T1-weighted images were preprocessed by the following steps. (1) Checking the BIDS JSON-format image meta-data; (2) intensity non-uniformity correction and skull-stripping; (3) tissue segmentation of cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM); (4) brain surface reconstruction; (5) deleting initial 10 time points; (6) boundary-based registration of BOLD and T1 images; (7) BOLD image spatial

¹ <https://docker.com>

normalization to fsaverage5 space; (8) head-motion, WM, and CSF signal and linear trend nuisance regression; (9) bandpass filtering (0.01–0.1 Hz); (10) spatial smoothing [full-width at half-maximum (FWHM) of 6 mm]. Detailed preprocessing procedures can be found in our previous research (Chen and Yan, 2021).

Of note, slice-timing corrections were not conducted because there were errors in the slice-timing information of some participants. Normally, DPABISurf/DPARSF would read the slice-timing information from DICOM header files (if the input images were in DICOM format) and metadata files in the BIDS format or the DPABI format (if the input images were in NIFTI format). As the demonstrating data in the fmri-open-qc-rest dataset were in NIFTI format, the slice-timing correction procedures would use the related metadata in the BIDS schema. The related information such as acquisition time for each slice and the scanning sequence (e.g., interleave or sequence while scanning different slices in a volume) were recorded in separated JSON files in the BIDS data-structure and could not be extracted from the NIFTI images themselves. In the fmri-open-qc-rest dataset, slice-timing-related information of some participants was missing or incorrect. The exact details were provided in see Section “3.2. Issues in MRI meta-data.” Therefore, we skipped the slice-timing correction while this procedure might be necessary for the images with a relatively long repetition time (Sladky et al., 2011) (e.g., TR = 2.5 for most of the participants in the dataset).

2.3. Volume-based MRI preprocessing

Volume-based data preprocessing in our study was carried out using the Data Processing Assistant for resting-state fMRI (DPARSF) (Yan and Zang, 2010), which was based on SPM (Friston et al., 1994) and had been integrated into Data Processing and Analysis of Brain Imaging (DPABI) (Yan et al., 2016). The first 10 time-points of the fMRI series were discarded. The head motion was corrected by a six-parameter (rigid body) linear transformation with a two-pass procedure (Yan et al., 2013). Reorient/QC was a module in DPARSF pipeline for both adjusting the orientation of the images and visually checking the image quality of each T1-weighted or BOLD image. We rated each image by a 5-point scale. The 5-point rating scales provided semiquantitative scores for the results of the visually evaluation in reorient/QC module. More points equaled better images. The derived reports would record both the rating scores and the comments for images. After the whole Reorient/QC procedures were finished, a QC-score-threshold of 3 was set in the following dialog box. The images with extremely bad quality were not be involved in the further preprocessing procedure to avoid contaminating other samples in the certain procedures (e.g., creating a group template). After coregistering the structural and functional images and unified segmentation (Ashburner and Friston, 2005) on T1 image, spatial normalization to MNI-152 space [a coordinate system created by Montreal Neurological Institute (Fonov et al., 2009)] was performed using the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) tool (Goto et al., 2013). The Friston 24-parameter model (Friston et al., 1996) was applied to regress out head motion effects. White matter signal, cerebrospinal fluid signal and linear trends were regressed out

from each voxel's time course. Finally, all images were filtered by temporal bandpass filtering (0.01–0.1 Hz) to reduce the effect of low-frequency drift and high-frequency physiological noise.

2.4. Quality control procedure

In general, we adopted six DPABI-derived reports to exclude participants with insufficient quality. The detailed criteria according to the reports were listed in Table 1. The QC procedures were integrated into two pipelines with graphic user interfaces (GUI) for the volume-based methods and surface-based methods. A detailed introduction to these modules could be found in the related course at <http://rfmri.org/Course>. An intuitive exclusive tool for checking spatial normalization quality in the volume-based preprocessing was displayed in Figure 1. The detailed criteria for eliminating samples derived from these reports were listed in Table 1.

- A. The QC rating scores derived from the Reorient/QC module in the DPARSF pipeline. The Reorient/QC module is a GUI designed for visually checking and manually orientation-adjusting the raw T1-weighted and functional images. The QC scores for each subject were given by the user according to the imaging quality. Subjects with structural or functional image QC scores below 3 would not be included in further preprocessing.
- B. The head-motion reports from DPABISurf/DPARSF pipeline. There were two reports about the head-motion of participants. The first one was a brief report for excluding participants according to several commonly-used rules (e.g., maximum rigid displacement or rotation exceeding 3 mm or 3 degrees). The second one was a detailed head-motion report spreadsheet recording the head-motion in different directions and the framewise displacements (FD) would be used as another threshold of mean head-motion (Jenkinson et al., 2002). We set a mean FD-Jenkinson head-motion threshold to 0.2.
- C. The dynamic graph for checking co-registration between structural images and functional images of each participant derived from DPABISurf pipeline. Bad BOLD-T1 co-registration, MRI artifacts and flipped image direction can be distinguished from this report.
- D. The dynamic graph for checking brain surface reconstruction for each participant derived from DPABISurf pipeline. Bad brain surface reconstruction can be distinguished from this report. Of note, bad skull stripping may lead to inaccurate surface reconstruction and structural metrics estimation and can be recognized in this report.
- E. The dynamic graph for checking spatial normalization from individual space to standard (MNI) space of each participant derived from DPABISurf pipeline. Bad spatial normalization, MRI artifacts, low signal-to-noise ratio, anomalous structural occupancy or abnormality can be distinguished from this report. The three graphical reports (e.g., co-registration, surface reconstruction and spatial normalization) of every participant were summarized into three HTML page in the derived QC folder in the DPABI working directory.

TABLE 1 Resting state functional magnetic resonance imaging (fMRI) quality control (QC) criteria: Exclude a subject if.

Index	Criteria	Derived from
A1	Low brain coverage (quantitative and qualitative)	DPARF, QC report
A2	Severe signal losses in temporal lobe (qualitative)	DPARF, QC report
A3	Head-motion related artifacts (qualitative)	DPARF, QC report
A4	Other MRI artifacts (qualitative)	DPARF, QC report
A5	Flipped/Uncertain scan direction (qualitative)	DPARF, QC report
A6	Anomalous structural occupancy or abnormality (qualitative)	DPARF, QC report
B1	Maximum head-motion exceeding 3 mm or 3 degree (quantitative)	DPARF/DPABISurf, Realign parameters
B2	Averaged framewise displacements exceeding 0.2 (quantitative)	DPARF/DPABISurf, Realign parameters
C1	Bad BOLD-T1 co-registration (qualitative)	DPABISurf, QC_EPItot1 report
C2	Head-motion related artifacts (qualitative)	DPABISurf, QC_EPItot1 report
C3	Other MRI artifacts (qualitative)	DPABISurf, QC_EPItot1 report
C4	Flipped/Uncertain scan direction (qualitative)	DPABISurf, QC_EPItot1 report
D1	Bad brain surface reconstruction (qualitative)	DPABISurf, QC_SurfaceReconstruction report
D2	Bad skull stripping (qualitative)	DPABISurf, QC_SurfaceReconstruction report
E1	Bad spatial normalization (qualitative)	DPABISurf, QC_T1toMNI report
E2	Head-motion-related artifacts (qualitative)	DPABISurf, QC_T1toMNI report
E3	Other MRI artifacts (qualitative)	DPABISurf, QC_T1toMNI report
E4	Low signal-to-noise ratio (qualitative)	DPABISurf, QC_T1toMNI report
E5	Anomalous structural occupancy or abnormality (qualitative)	DPABISurf, QC_T1toMNI report
F1	Abnormal TR, number of volumes, etc., (quantitative)	DPARF/DPABISurf, Meta-data report

“Other MRI artifacts” indicate a variety of visually recognizable MRI artifacts, including susceptibility artifacts, wraparound artifacts, coil-related artifacts, chemical artifacts, etc.

F. The meta-data report spreadsheet (TRInfo.tsv) of images generated by DPARF or DPABISurf. Abnormal meta-data records such as a smaller number of volumes, atypical TR and strange voxel sizes can be distinguished from this report. This report was considered a unique QC resource in DPABI because the mistakenly included images and incomplete images could be easily discriminated using the meta-data reports.

fMRI metrics included regional homogeneity (ReHo), (fractional) amplitude of low-frequency fluctuations (fALFF/ALFF) and degree centrality (DC). The sites and the mean FD-Jenkinson scores were included as covariates. The statistical maps of the two-sample *t*-tests were corrected for family-wise error rate (FWER) using Gaussian random field (GRF) correction. The vertex-wise threshold was 0.001 and the cluster-wise threshold for GRF correction was 0.017 (0.05/3, 3 for Bonferroni correction of two hemispheres and one subcortical area).

2.5. Sex difference with/without quality control

To preliminarily illustrate the effect of quality control in statistical analysis, we conducted two-sample *t*-tests to show the sex differences in some common fMRI metrics. Of note, a comprehensive evaluation of the QC-effect in group-level analysis (e.g., taking into account the site-effect and the reduced sample size after eliminating samples) would be a larger and separate topic. Importantly, the sex labels of the participants were not provided by the organizers of fmri-open-qc-rest dataset and we used a T1-weighted image-based classifier to predict the sex of each participant (Lu et al., 2022). Considering the sex classifier achieved about 95% accuracy, we supposed that the estimated classifier values would be close to the ground truth. Sex differences were tested in both the images with QC and the images without any QC. For the statistics without QC, thirteen estimated male participants and seven estimated female participants were excluded. The

3. Results

3.1. Quality control summary

In sum, 12 participants were excluded after quality control in DPABI and 8 participants might be further excluded on a stricter standard, accounting for 8.6 and 5.8% of the whole fmri-open-qc-rest sample (please see a detailed excluding list with subject ID in [Supplementary Table 1](#)). The detailed QC criteria were described in the following sections. The orders of these sections were determined by the frequency of being triggered and the importance of the excluding criterion in each section (e.g., from high to low).

3.2. Issues in MRI meta-data

There were several potential issues in the meta-data of images that were identified before preprocessing.

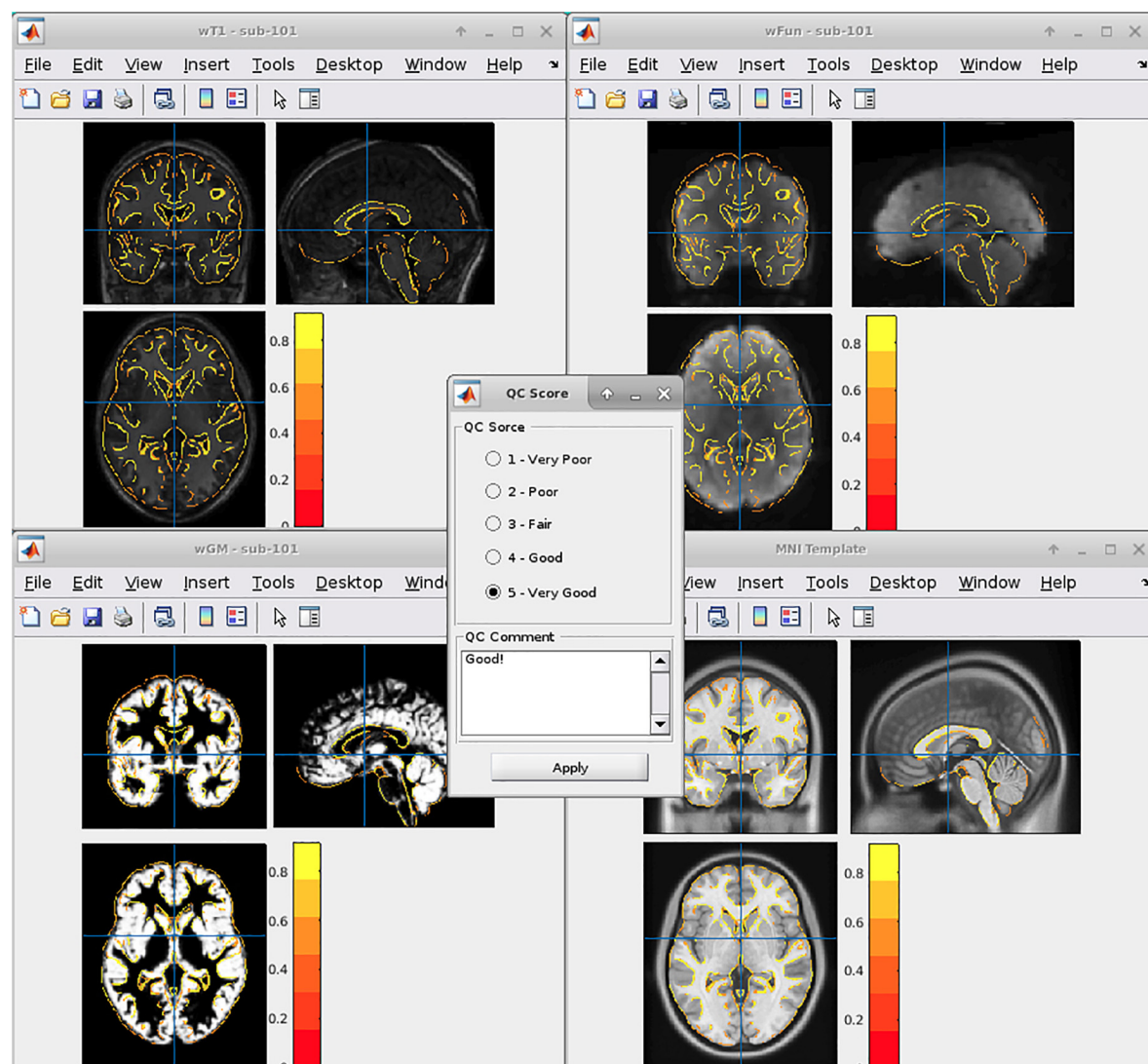


FIGURE 1
Graphic user interfaces of the spatial normalization quality control (QC) tools in DPABI.

Firstly, the functional images in site-2 and site-5 could not pass the BIDS metadata validation procedure in DPABI. The bids-validation tools reported that “slice-timing values contain invalid value as it is greater than the repetition time” for five participants (e.g., from sub-501 to sub-504, sub-509). Therefore, the five participants with the specific slice-timing errors were labeled as “uncertain,” as we suspected the acquisition sequences were thoroughly distorted. In addition, some of the participants did not have any slice-timing information in the BIDS schema. As we skipped the slice-timing correction in preprocessing, these participants were not excluded from the present study.

Secondly, the number of volumes (time points) was not consistent in site-1 and site-6. It may be acceptable for site-6 as we anticipate that site-6 were constructed by multiple sub-site. But the two participants (e.g., sub-114 and sub-115) with fewer volumes compared with the others in site-1 may suggest data loss in practice. We did not label these suspicious samples as “uncertain” or “excluded” as we did not know the actual scanning protocols for

these participants. However, we still raised this frequently occurring issue (inconsistent number of volumes for the images with the same scanning protocol) to inform the beginner of MRI data processing.

Thirdly, sub-605 had two runs of the BOLD series in the raw data while the others only had one run in each session. No additional information was available to help determine which run was more appropriate for further processing. We arbitrarily used the latter one and did not exclude this participant. Because in the practice, the additional run of an MRI series was probably due to the unsatisfying quality of the previous run of the same series (e.g., head-motion exceeding the criteria).

3.3. Head-motion related artifacts

The head-motion induced artifacts were the most frequently reported issue in the QC procedure. Seven out of twenty “uncertain” or “excluded” participants were potentially excluded

due to unacceptable head-motion. Some of them were visually identified and the others were identified by the head-motion report generated by DPARSF/DPABISurf (Figure 2A). Of note, the criteria related to the head-motion should be determined according to the research topic (Nebel et al., 2022).

3.4. Bad brain surface reconstruction

The core procedure of the surface-based methods was brain surface reconstruction. The surface reconstruction could fail due to a variety of quality problems (e.g., low brain coverage of field of view, low signal-to-noise ratio, abnormal brain structure and imaging artifacts, Figure 2B). In addition, the low quality of skull stripping may also hamper accurate surface reconstruction (Figure 2C).

3.5. Bad spatial normalization

There were two structural images of the participants that failed to achieve satisfying spatial normalization (Figure 2D). Spatial normalization (and related structural segmentation) could fail due to the low quality of images and local minimum in optimization induced by certain random seeds under extremely rare circumstances. Spatial normalization could be substantially improved by the reorientation procedure (e.g., manually rigid translation and rotation before spatial normalization) in DPARSF.

3.6. Other MRI artifacts

Besides head-motion, there are many MRI artifacts that could affect the image quality, including magnetic susceptibility artifacts, wraparound artifacts, coil-related artifacts, chemical artifacts and et al. (e.g., the T1-weighted images of sub-305 were blurred by unknown MRI artifacts, Figure 2E).

3.7. Abnormal brain structures

It's very challenging for neuroscientists to distinguish abnormal brain structures from normal anatomy or tiny MRI artifacts (Figures 2F, G). For example, sub-509 was labeled as uncertain because of the large ventricle. The QC classifiers of the UKBiobank would also take "Bad registration: Structurally atypical: Big Ventricles" as a problem situation. However, large ventricles might be common in the aged population and may not relate to disorders. Therefore, the eliminating criteria could be changeable according to the aim of the studies.

3.8. Flipped Z-axis direction

The functional MRIs of two subjects (sub-518 and sub-519) were flipped along the z-axis (Figure 2H). These results underlined the importance of visually checking the images. Flipped images along z-axis (up-down) could be further reversed and are

less destructive, but images flipping along the x-axis (left-right) would be harder to recognize and would significantly affect brain symmetry research.

3.9. Sex differences with/without quality control

As shown in Figure 3, both of the statistical maps of ReHo sex differences (with/without QC) showed significantly decreased spontaneous activity strength in the posterior cingulate cortex in the male group, which was consistent with the pre-existing literature (Chen et al., 2018). However, the maximal effect size values (Cohen's f^2) with QC (0.234 in the left hemisphere, 0.173 in the right hemisphere and 0.161 in the subcortical area) were higher than that without QC (0.221 in the left hemisphere, 0.152 in the right hemisphere and 0.153 in the subcortical area). Similarly, the maximal effect size values in the sex difference statistical maps of DC, fALFF, and ALFF with QC were higher than that without QC (Supplementary Figures 1–3).

4. Discussion

In the present study, a well-organized open-available MRI dataset was quality controlled by DPABI pipelines both in volume space and surface space. Twenty (14.4%) participants were categorized as excluded or uncertain. The reasons for these participants to be excluded could be summarized into eight categories: MRI meta-data issues, head-motion related artifacts, bad brain surface reconstruction, bad spatial normalization, other MRI artifacts, abnormal brain structures, and flipped images. In general, we believed that the QC procedure in DPABI could effectively improve the validity of the following analysis.

As mentioned in the description of fMRI Open QC Project, there is no single correct way to do QC. The criteria (thresholds) should be adjusted according to the population and the aim of the study. For example, head-motion related artifacts are still the most prevalent reason for excluding participants. Three types of criteria for controlling head-motion effect were used in the present study: (1) visual screening, (2) thresholding maximum head-motion, and (3) thresholding mean FD-Jenkinson. For studies whose research population is children or brain disorder patients, setting a strict threshold may dramatically reduce the available samples which is not acceptable for some longitudinal studies. While for studies in which head-motion artifacts must be minimized, some time-consuming but effective algorithms such as ICA-AROMA (Pruim et al., 2015) could be used to further remove head-motion effects. Another example is that participants with extremely large ventricles might be excluded from a group of children, but might be kept in a group of aged participants. In addition, all the QC criteria should be taken into account to determine the imaging quality of a participant. For example, the quality of skull stripping is low for both sub-312 and sub-315. But sub-312 was categorized as "uncertain" while sub-315 was categorized as "excluded" due to the additional uncertain structural occupancy and artifact on the parietal lobe. In addition, some of the QC procedures in DPABI were not conducted in the present study. For example,

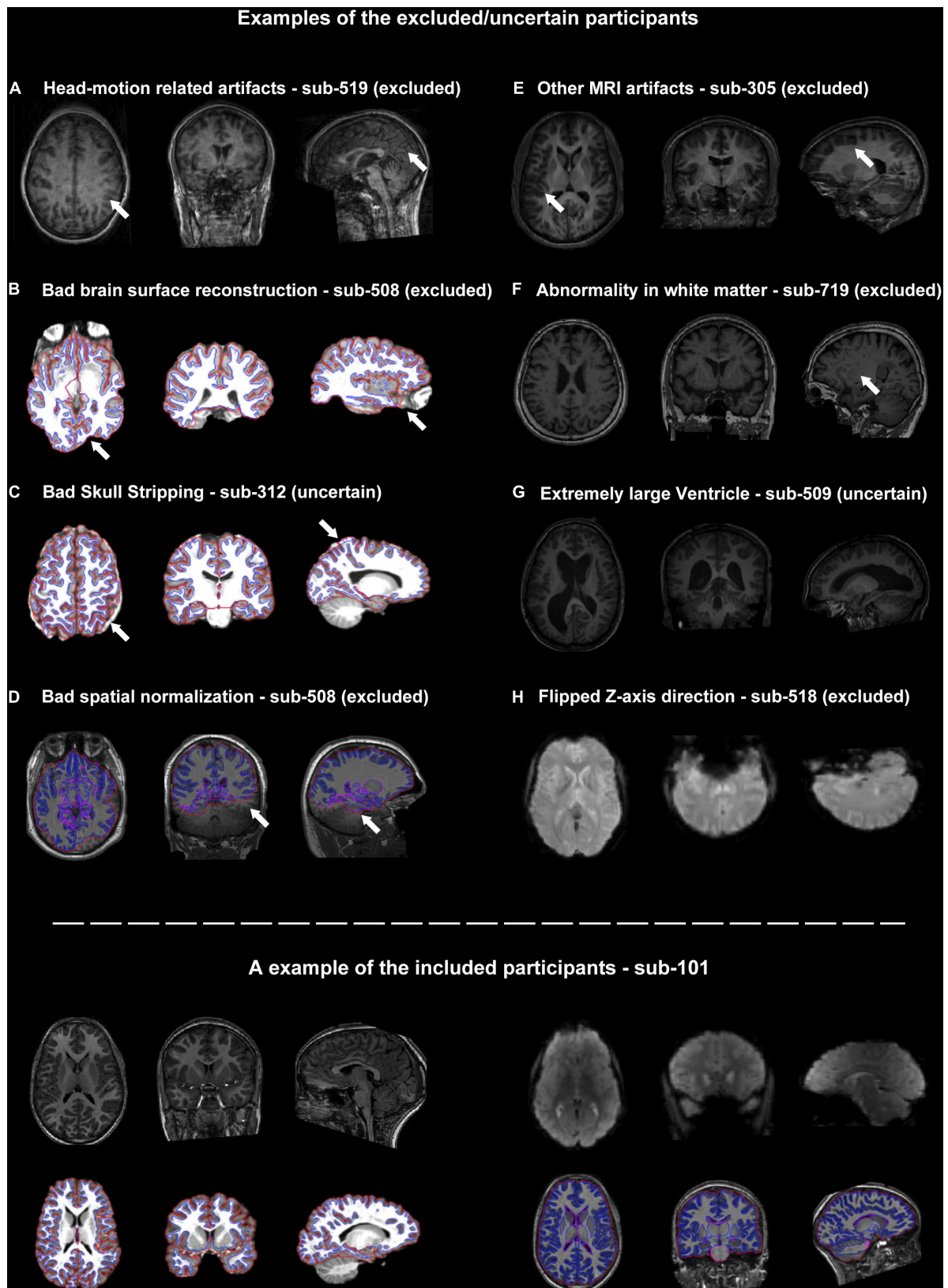
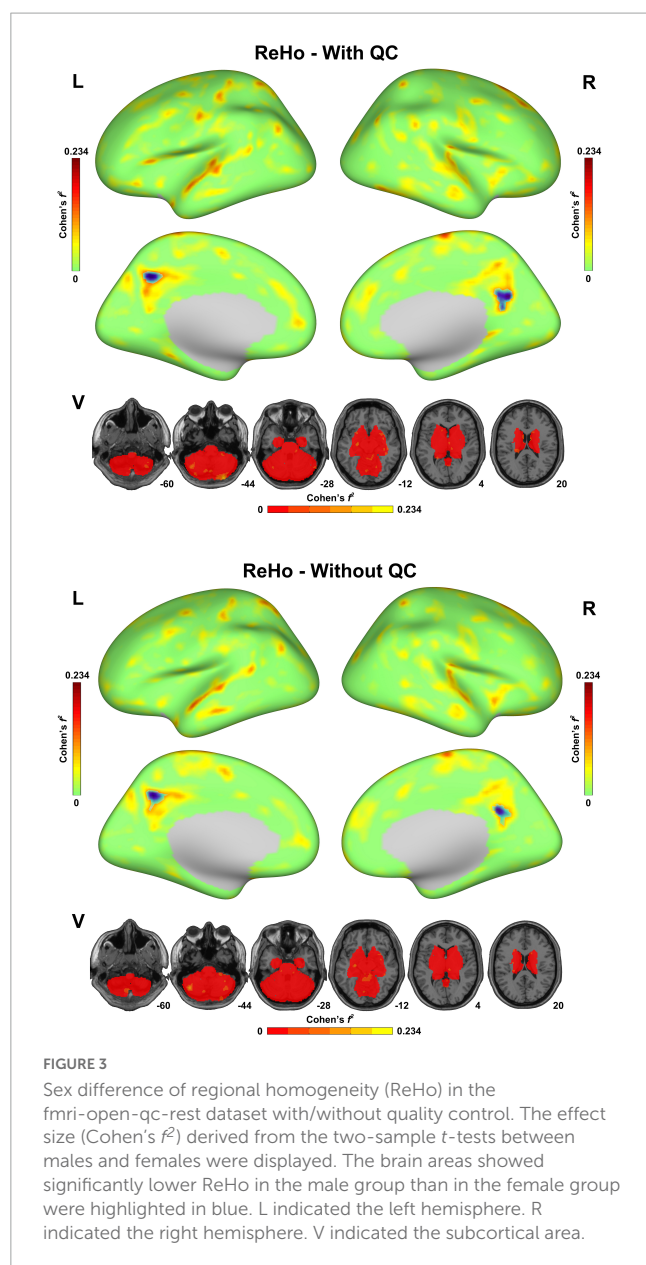


FIGURE 2

Representative examples of quality control (QC) items for which subjects were categorized as excluded or uncertain. (A–H) Examples of images with inadequate quality. The suspicious areas were highlighted using white arrows. The lower panel of the graph showed an example of the included participants.



ICA-AROMA is an outstanding algorithm to control head-motion related artifacts based on independent component analysis (ICA). As this algorithm is extremely time-consuming, it is an optional method in DPABISurf but is not conducted in default, while some other pipelines would include ICA-based nuisance regressions using a modified ICA-AROMA algorithm (Waller et al., 2022). Moreover, a detailed list of exclusion criteria and excluded subject IDs in the studies based on public datasets would save time for other researchers and improve the reproducibility of the findings.

Eliminating participants with bad image quality is a critical procedure to improve the quality of research. In a broader sense, the quality control in fMRI research should also include the daily scanner QC using water phantom, contraindications inspection (e.g., metal braces) while recruiting participants, correct patient positioning, head-motion suppression using sponge mat or optimized coil, avoiding meta-data loss at image archive platforms, checking critical meta-data before preprocessing,

carefully eliminating participants using QC reports generated by preprocessing pipelines, rigorous coding and statistic, etc. The acquisition protocols also interact with the QC procedure. For example, the multiband acquisition could improve the temporal resolution but decrease the signal-to-noise ratio (SNR) (Smith et al., 2013). Therefore, the SNR should be included as an important criterion in studies using multiband protocols. Discussing all these procedures is out of the scope of the present study, but the steps mentioned above would also influence participant-eliminating.

Therefore, more automatic QC tools are critical. For example, the sex of participants could be mistakenly recorded, and this mistake is hard to recognize. Recently, a T1-weighted image-based classifier trained using more than 85,000 MRI samples from more than 217 sites/scanners achieved 95% accuracy in a sex classification task on the independent datasets. This sex classifier could be an *Ex post* check procedure for sex labels.² As mentioned in the results 3.8 section, flipped images along the x-axis (left-right) would be a very subtle situation that is not easy to distinguish. The oil capsule marks for labeling left or right are not available for every dataset and the tricks [e.g., brain torque (Toga and Thompson, 2003)] for visual checks may not work for every participant. Fortunately, an efficient tool built in the AFNI fMRI processing procedure that can automatically distinguish the flipped images has been developed (Glen et al., 2020). Besides the specialized QC modules in DPABI, the input preparer module and the data organization checking module could also help avoid including incomplete images. And a new harmonization module in DPABI containing comprehensive multi-center imaging harmonizing methods would be available soon. In addition, as mentioned in the introduction, the design philosophy of DPABI was to minimize the repetitive and non-standardized human involvement in fMRI preprocessing, but the decision-making part of human involvement is still unavoidable. The UKBiobank imaging team has developed an automated machine learning based QC tool which performed excellently on the UKBiobank dataset. However, the UKBiobank's scanning protocols are uniform across all of the scanning sites, which might result in overfitting and poor generalizability. The generalizability of this promising tool needs to be further validated on a variety of datasets.

In summary, the QC procedures for fMRI in DPABI are illustrated by preprocessing a well-organized open dataset. A set of reports derived from DPABI pipelines could be utilized for excluding images with bad quality. More automatic QC tools are needed in the big-data era while visually inspecting images is still indispensable.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://osf.io/qaesm/wiki/home/>; https://github.com/Chaogan-Yan/PaperScripts/blob/master/Lu_2023_fMRIQC.

² <http://brainimagenet.org>

Ethics statement

The studies involving human participants were reviewed and approved by Institute of Psychology. The patients/participants provided their written informed consent to participate in this study.

Author contributions

C-GY designed the overall experiment and the QC tools. BL carried out the QC procedure. Both authors contributed to the article, wrote the manuscript, and approved the submitted version.

Funding

This work was supported by the Sci-Tech Innovation 2030–Major Project of Brain Science and Brain-inspired Intelligence Technology (Grant Number: 2021ZD0200600), National Key R&D Program of China (Grant Number: 2017YFC1309902), the National Natural Science Foundation of China (Grant Numbers: 82122035, 81671774, and 81630031), the 13th Five-year Informatization Plan of Chinese Academy of Sciences (Grant Number: XXH13505), the Key Research Program of the Chinese Academy of Sciences (Grant Number: ZDBS-SSW-JSC006), Beijing Nova Program of Science and Technology (Grant Number: Z191100001119104), and the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences (Grant Number: E2CX4425YZ).

References

- Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034
- Ashburner, J. (2012). SPM: A history. *Neuroimage* 62, 791–800. doi: 10.1016/j.neuroimage.2011.10.025
- Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *Neuroimage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018
- Backhausen, L. L., Herting, M. M., Buse, J., Roessner, V., Smolka, M. N., and Vetter, N. C. (2016). Quality control of structural mri images applied using freesurfer—a hands-on workflow to rate motion artifacts. *Front. Neurosci.* 10:558. doi: 10.3389/fnins.2016.00558
- Bastiani, M., Cottaar, M., Fitzgibbon, S. P., Suri, S., Alfaro-Almagro, F., Sotiropoulos, S. N., et al. (2019). Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction. *Neuroimage* 184, 801–812. doi: 10.1016/j.neuroimage.2018.09.073
- Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Chen, X., and Yan, C. G. (2021). Hypostability in the default mode network and hyperstability in the frontoparietal control network of dynamic functional architecture during rumination. *Neuroimage* 241:118427. doi: 10.1016/j.neuroimage.2021.118427
- Chen, X., Lu, B., and Yan, C. G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Hum. Brain Mapp.* 39, 300–318. doi: 10.1002/hbm.23843
- Coalson, T. S., Van Essen D. C., and Glasser, M. F. (2018). The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proc. Natl. Acad. Sci. U. S. A.* 115, E6356–E6365. doi: 10.1073/pnas.1801582115
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* 12:e0184661. doi: 10.1371/journal.pone.0184661
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Med.* 16, 111–116. doi: 10.1038/s41592-018-0235-4
- Fischl, B. (2012). FreeSurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021
- Fonov, V. S., Evans, A. C., McKinstry, R. C., Alml, C. R., and Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage* 47. doi: 10.1016/S1053-8119(09)70884-5
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312
- Glen, D. R., Taylor, P. A., Buchsbaum, B. R., Cox, R. W., and Reynolds, R. C. (2020). Beware (surprisingly common) left-right flips in your mri data: An efficient and robust method to check mri dataset consistency using AFNI. *Front. Neuroinform.* 14:18. doi: 10.3389/fninf.2020.00018

Acknowledgments

We would like to acknowledge the topic editors especially Dr. Paul Taylor for organizing the demonstrating quality control procedures in fMRI research topic and editing the present manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2023.1069639/full#supplementary-material>

- Goto, M., Abe, O., Aoki, S., Hayashi, N., Miyati, T., Takao, H., et al. (2013). Diffeomorphic anatomical registration through exponentiated lie algebra provides reduced effect of scanner for cortex volumetry with atlas-based method in healthy subjects. *Neuroradiology* 55, 869–875. doi: 10.1007/s00234-013-1193-2
- Hagler, D. J. Jr., Hattton, S., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., et al. (2019). Image processing and analysis methods for the adolescent brain cognitive development study. *Neuroimage* 202:116091.
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* 264, 47–56. doi: 10.1016/j.jneumeth.2016.03.001
- Lu, B., Li, H. X., Chang, Z. K., Li, L., Chen, N. X., Zhu, Z. C., et al. (2022). A practical Alzheimer's disease classifier via brain imaging-based deep learning on 85,721 samples. *J. Big Data* 9:101. doi: 10.1186/s40537-022-00650-y
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *Elife* 10:e71774. doi: 10.7554/eLife.71774
- Michael, M., Francisco, C., Adriana, D. M., Clare, K., Maarten, M., Stanley, C., et al. (2013). Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC). *Front. Neuroinform.* 7. doi: 10.3389/conf.fninf.2013.09.00042
- Mortamet, B., Bernstein, M. A., Jack, C. R. Jr., Gunter, J. L., Ward, C., Britson, P. J., et al. (2009). Automatic quality assessment in structural brain magnetic resonance imaging. *Magn. Reson. Med.* 62, 365–372. doi: 10.1002/mrm.21992
- Nebel, M. B., Lidstone, D. E., Wang, L., Benkeser, D., Mostofsky, S. H., and Risk, B. B. (2022). Accounting for motion in resting-state fMRI: What part of the spectrum are we characterizing in autism spectrum disorder? *Neuroimage* 257:119296. doi: 10.1016/j.neuroimage.2022.119296
- Pruim, R. H. R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., and Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage* 112, 267–277. doi: 10.1016/j.neuroimage.2015.02.064
- Sladky, R., Friston, K. J., Trostl, J., Cunnington, R., Moser, E., and Windischberger, C. (2011). Slice-timing effects and their correction in functional MRI. *Neuroimage* 58, 588–594. doi: 10.1016/j.neuroimage.2011.06.078
- Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., et al. (2013). Resting-state fMRI in the human connectome project. *Neuroimage* 80, 144–168. doi: 10.1016/j.neuroimage.2013.05.039
- Tange, O. (2011). Gnu parallel-the command-line power tool. *USENIX Mag.* 36, 42–47.
- Toga, A. W., and Thompson, P. M. (2003). Mapping brain asymmetry. *Nat. Rev. Neurosci.* 4, 37–48. doi: 10.1038/nrn1009
- Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., et al. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage* 99, 166–179. doi: 10.1016/j.neuroimage.2014.05.044
- Waller, L., Erk, S., Pozzi, E., Toenders, Y. J., Haswell, C. C., Buttner, M., et al. (2022). ENIGMA HALPipe: Interactive, reproducible, and efficient analysis for resting-state and task-based fMRI data. *Hum. Brain Mapp.* 43, 2727–2742. doi: 10.1002/hbm.25829
- Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage* 92, 381–397. doi: 10.1016/j.neuroimage.2014.01.060
- Yan, C. G., and Zang, Y. F. (2010). DPARSF: A MATLAB toolbox for “pipeline” data analysis of resting-state fMRI. *Front. Syst. Neurosci.* 4:13. doi: 10.3389/fnsys.2010.00013
- Yan, C. G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di, M. A., et al. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *Neuroimage* 76, 183–201. doi: 10.1016/j.neuroimage.2013.03.004
- Yan, C. G., Wang, X. D., and Lu, B. (2021). DPABISurf: Data processing & analysis for brain imaging on surface. *Sci. Bull.* 66, 2453–2455. doi: 10.1016/j.scib.2021.09.016
- Yan, C. G., Wang, X. D., Zuo, X. N., and Zang, Y. F. (2016). DPABI: Data processing & analysis for (resting-state) brain imaging. *Neuroinformatics* 14, 339–351. doi: 10.1007/s12021-016-9299-4



OPEN ACCESS

EDITED BY

Paul A. Taylor,
National Institute of Mental Health (NIH),
United States

REVIEWED BY

Stephen J. Gotts,
National Institute of Mental Health (NIH),
United States
Suril Gohel,
Rutgers, The State University of New Jersey,
United States

*CORRESPONDENCE

Rasmus M. Birn
✉ rbirn@wisc.edu

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroimaging

RECEIVED 18 October 2022

ACCEPTED 17 February 2023

PUBLISHED 13 March 2023

CITATION

Birn RM (2023) Quality control procedures and
metrics for resting-state functional MRI.
Front. Neuroimaging 2:1072927.
doi: 10.3389/fnimg.2023.1072927

COPYRIGHT

© 2023 Birn. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Quality control procedures and metrics for resting-state functional MRI

Rasmus M. Birn^{1,2*}

¹Department of Psychiatry, University of Wisconsin-Madison, Madison, WI, United States, ²Department of Medical Physics, University of Wisconsin-Madison, Madison, WI, United States

The monitoring and assessment of data quality is an essential step in the acquisition and analysis of functional MRI (fMRI) data. Ideally data quality monitoring is performed while the data are being acquired and the subject is still in the MRI scanner so that any errors can be caught early and addressed. It is also important to perform data quality assessments at multiple points in the processing pipeline. This is particularly true when analyzing datasets with large numbers of subjects, coming from multiple investigators and/or institutions. These quality control procedures should monitor not only the quality of the original and processed data, but also the accuracy and consistency of acquisition parameters. Between-site differences in acquisition parameters can guide the choice of certain processing steps (e.g., resampling from oblique orientations, spatial smoothing). Various quality control metrics can determine what subjects to exclude from the group analyses, and can also guide additional processing steps that may be necessary. This paper describes a combination of qualitative and quantitative assessments to determine the quality of fMRI data. Processing is performed using the AFNI data analysis package. Qualitative assessments include visual inspection of the structural T1-weighted and fMRI echo-planar images, functional connectivity maps, functional connectivity strength, and temporal signal-to-noise maps concatenated from all subjects into a movie format. Quantitative metrics include the acquisition parameters, statistics about the level of subject motion, temporal signal-to-noise ratio, smoothness of the data, and the average functional connectivity strength. These measures are evaluated at different steps in the processing pipeline to catch gross abnormalities in the data, and to determine deviations in acquisition parameters, the alignment to template space, the level of head motion, and other sources of noise. We also evaluate the effect of different quantitative QC cutoffs, specifically the motion censoring threshold, and the impact of bandpass filtering. These qualitative and quantitative metrics can then provide information about what subjects to exclude and what subjects to examine more closely in the analysis of large datasets.

KEYWORDS

connectivity, motion, fMRI, artifacts, quality control

Introduction

Functional MRI (fMRI) signal changes are relatively small and sensitive to various sources of noise, such as scanner artifacts, head motion, and other physiological fluctuations. Generating functional activation or connectivity maps from the acquired data therefore typically consists of a number of processing steps aimed at reducing this noise and aligning the brain images into a common space for group-level analyses. The programs used to

perform this processing can vary between research groups, and each step often has multiple options that can be chosen by the researcher. An integral part of this processing pipeline is quality control (QC) to determine what processing steps or options are needed, to determine the source of any problems in the pipeline, to determine whether a subject should be excluded from group-level analyses, and ultimately to ensure the accuracy and validity of the final results.

Quality control should ideally be performed first in real-time, while the subject is being scanned and still in the MRI scanner. The advantage to this is that corrupted data can be immediately identified and then re-acquired or otherwise addressed. It is also critical to perform QC at multiple stages during the pre-processing. This QC can be both qualitative and quantitative. Qualitative measures, such as viewing the data at different stages during the processing, is extremely useful because of the myriad ways that the data can be corrupted or that the processing can go awry. A trained researcher can then determine what additional processing steps may be needed or what options or parameters should be adjusted. Quantitative measures of QC, such as the signal-to-noise ratio or the amount of head motion, are also important, particularly for large datasets where qualitative QC can be time consuming. These quantitative measures also allow for more reproducible analyses and inform the level of confidence in the final imaging results.

The most common problems affecting the quality of resting-state functional MRI data include imaging artifacts, subject head motion, and errors in aligning the data to a common template space. Imaging artifacts can include B0-field distortions or malfunctions in the RF coil leading to spikes or variations of signal intensity near malfunctioning coil elements. Head motion is common in fMRI and has long been recognized as a problem that needs to be minimized and reduced (Friston et al., 1996). Resting-state functional connectivity studies are particularly sensitive to the effects of motion since connectivity is measured by the temporal similarity of fMRI time series between two or more regions using some metric, such as the Pearson's correlation coefficient (Biswal et al., 1995). Two regions with correlated non-neuronal signal variations (noise) would result in an erroneously inflated functional connectivity, while two regions with uncorrelated noise would result in reduced connectivity. Even small amounts of motion can have significant impact on functional connectivity (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). Alignment of the functional data requires both the alignment of the T2*-weighted EPI to the T1-weighted structural image and the alignment of the T1-weighted structural to the template. The alignment between the EPI and T1 needs to take into account the differences in contrast between a T1-weighted and a T2*-weighted image. Alignment of the T1 to template space can involve non-linear transformations (e.g., image warping), and the accuracy of these depends of the quality of the removal of non-brain tissue ("brain extraction" or "skull-stripping"). Finally, problems can occur due to user error, such as prescribing an imaging volume that misses part of the brain or making an error in converting between file formats.

This paper provides several suggested QC procedures and measures for the analysis of resting-state functional MRI. This QC consists of both qualitative and quantitative measures, which are described in detail in the Methods section, and are applied to T1-weighted structural and resting-state functional MRI data

provided by the OpenQC project. Finally, a determination is made whether to include or exclude each participant from further analyses, or when the inclusion or exclusion is borderline or depends on other factors.

Methods

MRI data

The MRI data consisted of T1-weighted structural MRI scans and T2*-weighted echo-planar imaging (EPI) resting-state functional MRI scans from 139 subjects drawn from 7 different sites, provided by the OpenQC project. These data were drawn from various publicly available MRI data repositories—ABIDE, ABIDE-II (Di Martino et al., 2014), Functional Connectome Project (Biswal et al., 2010), and OpenNeuro (Markiewicz et al., 2021). The EPI datasets all had a single echo time and did not have simultaneous multi-slice acquisitions. B0-field inhomogeneity measures (e.g., B0-field maps or EPIs with reversed phase encoding) were not provided.

Processing pipeline

All data processing was performed using AFNI unless otherwise indicated (Cox, 1996). Processing scripts are available on GitHub: <https://github.com/rbirn/OpenQC>. The ICBM 152 non-linear atlas version 2009 was used as the template "MNI" brain (Fonov et al., 2011). The T1-weighted image volume was aligned to the MNI template by removing non-brain tissue signals and non-linearly warping the image to the template (using AFNI's *@SSwarper*). The T1-weighted image was segmented into gray matter, white matter, and CSF using FSL's *fast* (Zhang et al., 2001). The functional MRI echo-planar imaging (EPI) data were processed by first removing the first 4 volumes to assure that the magnetization is at steady-state. The data were then corrected for slice-timing differences (*3dTshift*), rotated and resampled to remove any oblique orientation (*3dWarp*), and registered to the first volume in each time series to reduce the effects of head motion (*3dvolreg*). B1-field inhomogeneities (bias field) were estimated using *N4BiasFieldCorrection* from ANTs (Tustison et al., 2010). The data were then divided by this bias field to correct for B1-field inhomogeneity. The echo-planar image was aligned to the T1-weighted structural scan using a 12-parameter affine transformation (*align_epi_anat.py*). The EPI-to-T1 and T1-to-template transformations were then concatenated and used to non-linearly warp the fMRI data to the MNI template. In order to further reduce the effects of physiological noise and head motion, several nuisance regressors were included in a general linear model and projected out of the data (*3dTproject*). These included the average signal over the whole brain, the average signal over eroded white matter, average signal over CSF, the 6 realignment parameters, and the temporal derivatives of each of these regressors. This general linear model also included 2 polynomials (to account for slow drifts) and a set of sines and cosines to band-pass filter the data from 0.01 to 0.1 Hz. Time points where the volume-to-volume motion exceeded a predefined motion censoring threshold, as well

as the preceding time points, were excluded (censored) from the nuisance regression. Three different motion censoring thresholds were evaluated: 0.2, 0.4, 1.0 mm. Prior studies have shown that one source of variability in multi-site studies are differences in the spatial smoothness of the data (Friedman et al., 2008). Since the data in this study were acquired at different sites and different spatial resolutions, rather than applying a fixed amount of spatial smoothing, the data were then iteratively smoothed to achieve a final full-width at half maximum (FWHM) of 8 mm (using *3dBlurToFWHM*). For comparison, the data processing was repeated without regressing out the average whole-brain signal (global signal regression, GSR).

Functional connectivity maps were generated for 4 seed regions of interest—4 mm radius spheres located in the posterior cingulate (MNI coordinate: 0, 50, 31), the left primary motor cortex (MNI coordinate: 36, 20, 60), left auditory cortex (MNI coordinate: 43, 25, 14), and the left primary visual cortex (MNI coordinate: 30, 87, 9). These seed regions identify the default mode network, motor network, auditory network, and visual network, respectively. The preprocessed signal was averaged over each seed region of interest, and the Pearson's correlation coefficient between this seed time course and all other voxel time courses was computed. In addition to these voxel-wise functional connectivity maps, connections between multiple regions across the whole brain was investigated by computing a functional connectivity matrix. The brain was divided into 333 regions of interest according to a parcellation by Gordon et al. (2016). The preprocessed signal was averaged over each region of interest, and all pairwise correlations were computed.

For comparison of QC metrics, data were also processed using the more automated pipeline provided with AFNI, *afni_proc.py*. This pipeline used as input the original resting-state EPI and the T1 processed (brain extracted and aligned to template space) by @SSwarper, and included the following processing steps: removal of first 4 time points; alignment of EPI to T1; volume registration (motion correction); non-linear warping to template space; nuisance regression using average signal over eroded white matter and CSF, motion, and their derivatives; band-pass filtering (0.01–0.1 Hz); and blurring to a FWHM of 8 mm. This pipeline by default derives a set of quality control metrics from each subject and assembles them into an html-formatted document that can be viewed in a web browser.

Quality control procedures

First, several imaging parameters were extracted from the data and compared. This included the spatial resolution (voxel size), matrix size, repetition time (TR), obliquity, and number of time points (image volumes) acquired. These values informed some of the processing choices and QC criteria. Specifically, the fact that data were acquired at different spatial resolutions motivated iterative blurring of the data to a final resolution rather than applying a fixed spatial blur across subjects. The observation that some data were acquired with oblique orientations necessitated that this be accounted for when registering the EPI to the T1-weighted structural scan and the T1-weighted structural to the template. The total number of time points acquired needs to

be considered when applying certain QC criteria (e.g., the total number of “good” time points). The imaging parameters were also examined for any deviations from other scans acquired at that site. The processing pipeline described above was then run on each dataset. Log files were generated that contained any status or error messages (typically output to the screen). These log files were examined when the processing pipeline failed to produce the final preprocessed data output.

The image quality and alignment of each subject's T1-weighted structural scan to template space was examined by concatenating the T1-weighted images across subjects. Similarly, a single echo-planar image volume, after warping to template space but before nuisance regression or spatial smoothing, was extracted from each subject and concatenated across subjects. These series of image volumes were then be played as a movie within the AFNI GUI to identify any misalignments or other imaging artifacts. Functional connectivity maps for each of the seed regions were similarly concatenated and played as a movie, with the subject's T1-weighted image as the underlay and the functional connectivity as an overlay.

QC metrics

A number of quantitative metrics were computed, using the first (non *afni_proc.py*) pipeline described above, to assess data quality. These are briefly described below.

Left-right flip

Potential errors in the left-right orientation (i.e., accidental flips of the data in the L-R direction) were investigated by flipping the structural T1 dataset in the left-right direction and repeating the alignment between the EPI and T1. This is performed using the `-check_flip` option in AFNI's *align_epi_anat.py*. If the cost function for the alignment is lower for the flipped dataset, either the EPI or T1 is likely flipped in the L-R direction.

FWHM

The smoothness of the acquired EPI data were determined by computing the full-width at half-maximum (FWHM) in each of the 3 cardinal directions (using *3dFWHMx*). This measure can be used to determine whether variations in the image matrix are due to differences in the acquisition (e.g., acquiring data at a higher resolution) or to differences in the processing (e.g., resampling the data to a higher resolution). This information can then guide other processing choices, such as the amount of smoothing to apply, or whether to smooth to a predetermined amount of smoothness.

Temporal signal-to-noise ratio (TSNR)

The temporal signal to noise ratio was computed by dividing the mean signal over time in each voxel of the original acquired image by its standard deviation over time. This measure can be good at identifying data severely corrupted by head motion, RF coil problems (e.g., spiking), or other imaging artifacts. This measure does vary with the imaging parameters (resolution, number of averages, parallel imaging acceleration, field strength, echo time,

etc.), so it is difficult to set a strict cutoff. However, the average TSNR over the whole brain can be compared to other subjects within the group acquired with similar imaging parameters at that site.

Mean Enorm

Volume-to-volume head motion was assessed by first computing the temporal difference of each image realignment parameter (3 translations, 3 rotations), and then computing the Euclidean norm (square-root of the sum of squares, Enorm) of these temporal differences at each time point, with shifts in millimeters and rotations in degrees. Note that a 1 degree rotations corresponds to a displacement of 1 mm at a radius of 57 mm, roughly the distance from the center of mass to the edge of the brain. The mean value of the Enorm across time provides a measure of the mean (average) volume-to-volume motion for that imaging run.

Max Enorm

The maximum of the Enorm time course (computed as described above) across time provides a measure of the maximum motion from one time point to the next. The rationale for excluding subjects based on the maximum motion is that large motion is more likely to be associated with changes in B0-field distortions, moving into different parts of the RF coil sensitivity, and spin-history effects. However, if large motion is infrequent, there are approaches to reduce the resultant signal changes (Birn et al., 2022).

Number of “good” time points

The number of time points remaining after censoring time points exceeding a certain motion (Enorm) threshold. A related, and from a quality control viewpoint equivalent, metric is the degrees-of-freedom remaining after censoring, band-pass filtering, and nuisance regression. Enough degrees-of-freedom should remain to accurately estimate the functional connectivity. A degree-of-freedom cutoff of 15 was used for this study. Studies have also shown that the specificity (Van Dijk et al., 2010), test-retest reliability (Birn et al., 2013) and the identification accuracy (Finn et al., 2015) of functional connectivity increases with both greater number of time points and duration of acquisition. A QC cutoff of at least 5 min of good data has been used by prior studies (Van Dijk et al., 2010; Power et al., 2014, 2015). However, 3 of the sites in this study acquired only 5 min of data or less. Therefore, a QC cutoff of 4 min was used for this study.

Dice_e2a

The Sorensen-Dice coefficient between the echo-planar fMRI brain image and T1-weighted structural is computed as two times the intersection between whole-brain masks of the echo-planar image and T1-weighted image (after alignment, in template space) divided by the sum of the areas of each of these masks. The goal of this metric is to measure the accuracy of the EPI-to-T1 alignment. This measure can be computed using the AFNI program *3ddot*.

Dice_a2t

The Sorensen-Dice coefficient between the T1-weighted structural and MNI template is computed similar as above, but with whole-brain masks of the T1-weighted and MNI template images. The goal of this metric is to measure the accuracy of the T1-to-template alignment.

FCS

The functional connectivity strength (FCS) is the average functional connectivity from each voxel to all other voxels in the brain. Mathematically this is identical to computing the correlation between each voxel time series and a scaled version of the global signal. This scaled version of the global signal is computed by dividing each voxel's signal intensity time course by its standard deviation over time, and then computing the average of these scaled signals over the entire brain. This metric can be used to identify abnormally high correlations that may result from some RF coil problems, for example a loose connection in one of the coil elements causing spikes in the signal. These signal spikes occur at the same time across large portions of the image thus causing the time courses to be highly correlated. The rationale for using this measure in addition to TSNR is that a single spike may not affect the TSNR very much, but can affect the correlation of that voxel time course with all other voxel in that slice.

Similarity to mean FC

The similarity of the mean functional connectivity is determined by computing the correlation between each subject's functional connectivity matrix and the group average functional connectivity matrix (using AFNI's *3ddot*). This metric can identify potential outliers in functional connectivity. For comparison, the similarity was also using the Euclidean distance between each subject's functional connectivity matrix and the group mean functional connectivity matrix. To distinguish this metric from the similarity using Pearson's correlation, we call this the “Dissimilarity” since a greater Euclidean distance is associated with a reduced similarity and thus greater dissimilarity. This was computed using AFNI's *3dcalc* and *3dROIstats*.

Determination of QC criteria

A common QC criterion is to exclude time points whose framewise displacement (volume-to-volume motion) exceeds 0.2 mm (Power et al., 2014, 2015). We wanted to examine whether this censoring threshold was appropriate for the current study. Therefore, the processing pipeline was run for 3 different motion censoring thresholds: 0.2, 0.4, and 1.0 mm. In addition, we compared the functional connectivity both with and without bandpass filtering.

One measure that has been used to assess the effectiveness of different processing choices is the correlation between the functional connectivity and a quality control metric, such as the mean Enorm—a measure referred to as QC-FC (Ciric et al., 2018). This is essentially testing whether there is a difference

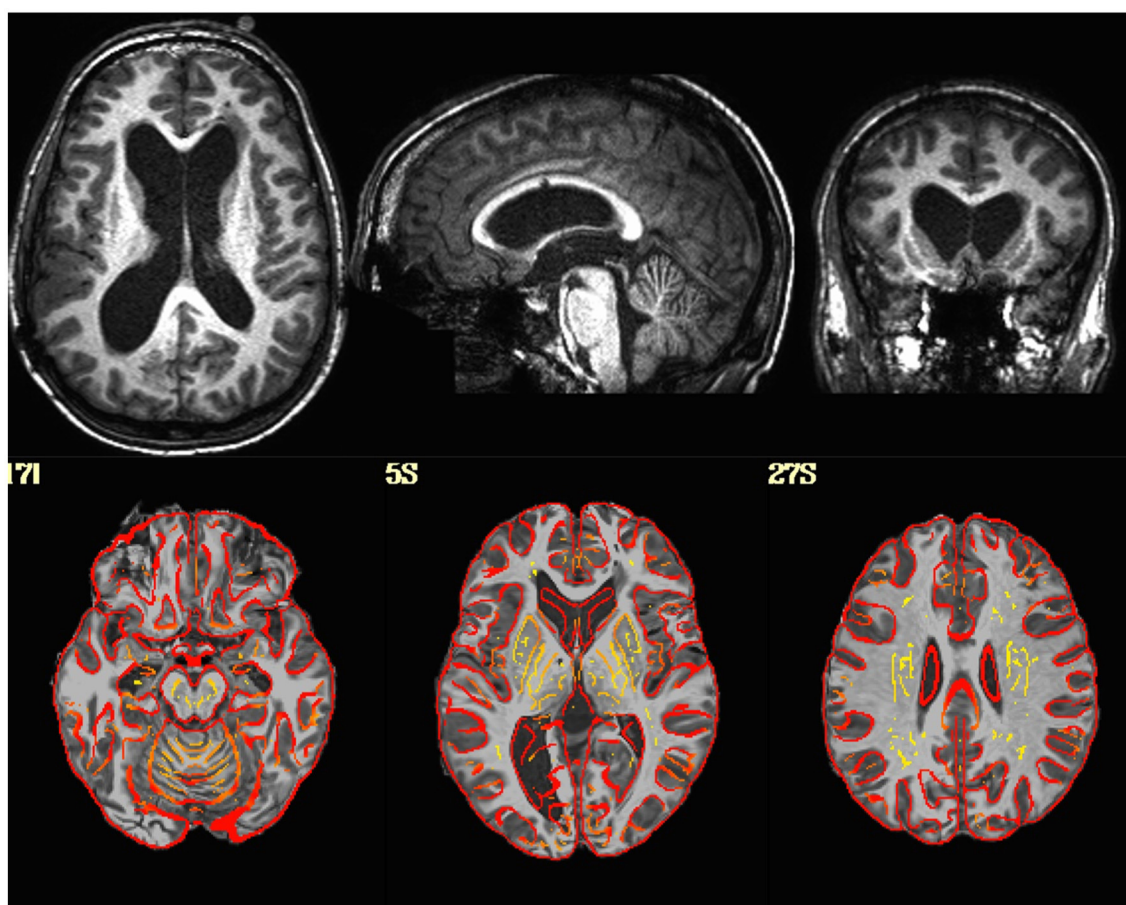


FIGURE 1

(Top Row) T1-weighted image in native space for a subject with enlarged ventricles. (Bottom Row) T1-weighted image non-linearly aligned to template space as underlay, with the gray/white matter boundaries from the template brain overlaid in red. Dice coefficient between the subject's T1 and the template = 0.96.

in functional connectivity as a function of head motion, i.e., between high-motion and low-motion subjects. We therefore computed the correlation between the functional connectivity and the mean Enorm for each connection in the connectivity matrix. We then computed a histogram of these correlation values. An additional metric that has been used to evaluate the effectiveness of different processing choices is the distance dependence of motion artifacts (Power et al., 2012, 2014, 2015; Ciric et al., 2018). This is computed as the correlation between the QC-FC metric described above and the distance between each of the nodes in the connectivity matrix.

We also looked at the similarity of each subject's functional connectivity matrix to the group average functional connectivity matrix, as described above. We then examined the correlation of this similarity with motion, specifically the mean Enorm. The rationale for the motion censoring threshold that we chose is provided in the results section (below).

Resources

The following software packages and versions were used in the analysis:

AFNI Version AFNI_21.2.07 (precompiled binary linux_openmp_64, Sep. 20, 2021).
FSL Version 6.0.4.
ANTs Version 0.0.0 (compiled May 26, 2020).

Results

The set of quality control (QC) summary criteria used for excluding or identifying problematic subjects in this study are shown in Table 1. The quality control procedures identified a number of problems with the data, leading to the exclusion of some of the subjects and modified processing for others. Very similar results were obtained from the *afni_proc.py* and our custom AFNI pipeline.

Examination of the imaging parameters showed that some of the datasets were acquired (or reconstructed) at a different matrix size compared to others from the same site. For example, sub-118 had a matrix size of 112 voxels while all other scans from that site had a matrix size of 96 voxels. The json files associated with the data all indicate that the data from this site was acquired with a matrix size of 84×81 . For site 5, 15 subjects had a matrix size of 80

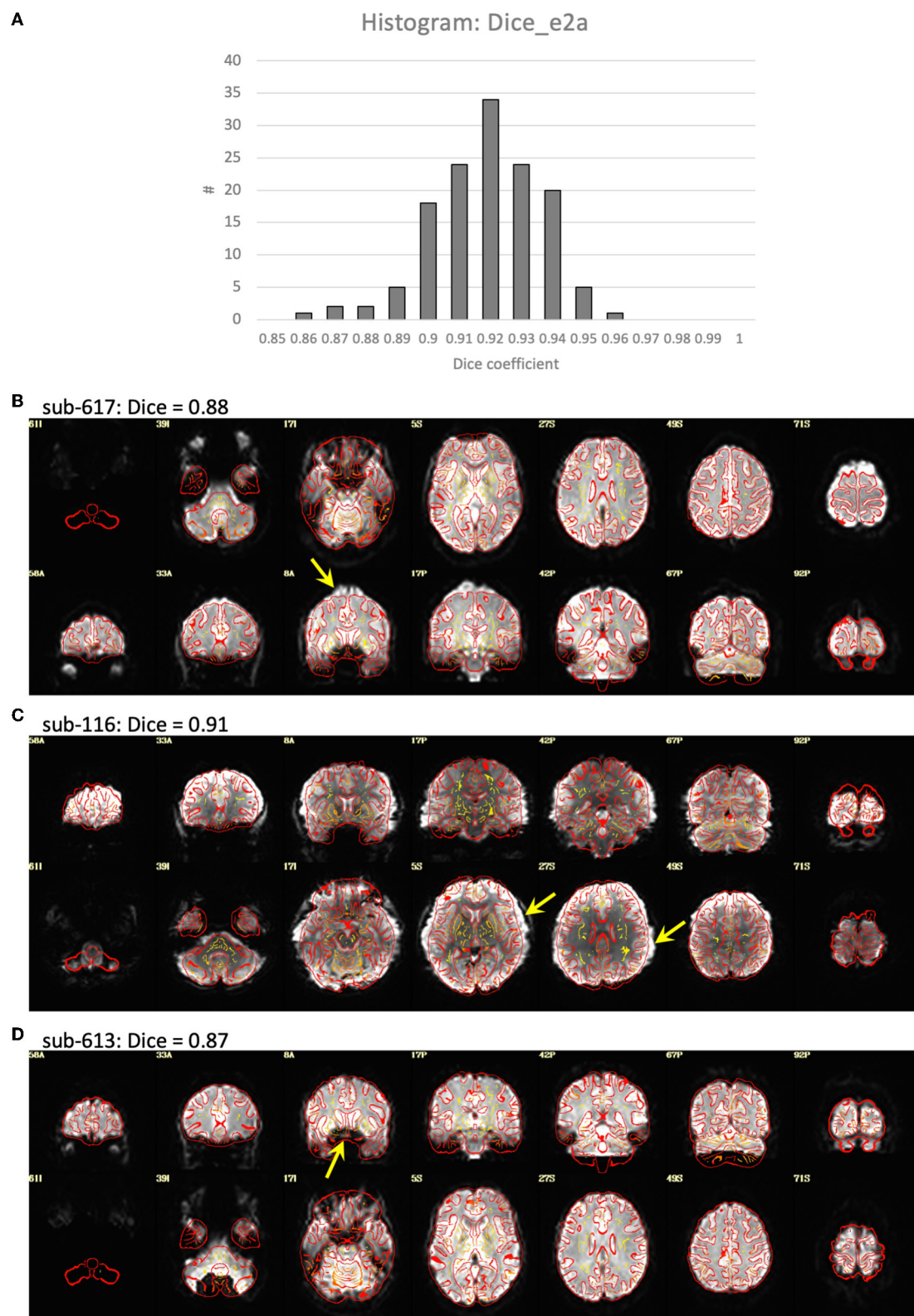


FIGURE 2

Alignment between the EPI and T1-weighted structural image. **(A)** Histogram of the Dice coefficients of the EPI and anatomical T1-weighted brain masks (Dice_e2a). **(B–D)** Case examples of the alignment between the EPI (in grayscale) and T1 (in red outline). **(B)** Subject 617 shows a slight misalignment between the EPI and T1 in the superior region of the brain (yellow arrow), and has a relatively low Dice coefficient = 0.88 compared to the rest of the group. **(C)** Subject 116 shows a slight misalignment, a stretching of the EPI in the left-right direction (yellow arrows), but has a Dice coefficient close to the mean of the group, Dice = 0.91. **(D)** Subject 613 shows a good alignment between the EPI and T1 in the cortex, but has a signal dropout in the frontal lobe resulting in a relatively low Dice coefficient = 0.87.

voxels while 5 subjects had a matrix size of 128 voxels. The datasets from this site with 128 voxels had significantly greater smoothness (FWHM in the x- and y-directions) compared to the datasets with 80 voxels ($p < 0.004$), suggesting that the data was re-interpolated after acquisition, resulting in increased blurring.

Visualization of the original EPI datasets indicated that two datasets (sub-518, sub-519) were upside down, with the I-S axis inverted. Alignment between the EPI and T1 indicated that two subjects (sub-101, sub-115) had either the EPI or T1 flipped in the L-R direction. Visualization of the T1-weighted structural images indicated that one subject (sub-509) had much larger ventricles than the rest of the sample (Figure 1).

Visualization of the T1-weighted images concatenated across subjects and played as a movie indicated good alignment of each T1 to the template. Alignment of the EPI to template space was generally quite good, but had a greater variability across subjects with some brain areas showing a slight misalignment to the template brain in some subjects (Figure 2). Closer examination of the processing in these subjects indicated that this misalignment to template space was due to a poor alignment between the EPI and T1-weighted image, even after automatic alignment. The Dice coefficient between the EPI and T1 (Dice_e2a) was lower for some of the misaligned participants compared to the rest of the group. However, some participants had lower Dice coefficients due to B0-field inhomogeneity induced signal dropout, and other subjects had Dice coefficients close to the group mean despite showing substantial misalignments (Figure 2).

As expected, temporal signal-to-noise ratio (TSNR) was reduced in subjects with higher amounts of motion (Figure 3). The converse was not necessarily true—some subjects with low motion also had low TSNR, possibly due to other non-motion sources of noise. No outliers or abnormalities were found in the temporal SNR or functional connectivity strength to indicate any coil artifacts. Similarly, the entire cortex was scanned in all subjects.

The most common problem across datasets was excessive head motion. At an Enorm censoring threshold of 0.2 mm, 15 subjects did not have enough degrees of freedom left for the nuisance regression and bandpass filtering. A total of 26 subjects had very low degrees of freedom (<15), and 16 subjects had <4 min of data left after censoring. At a censoring threshold of 0.4 mm, 2 subjects did not have enough degrees of freedom after censoring, 4 subjects had very low degrees of freedom, and 2 subjects had <4 min of data left after censoring. Two subjects had one or more movements >3 mm. A closer examination of the subject with the largest motion of 6.5 mm (sub-102) revealed that the motion occurred right at the end of the imaging run (Figure 4). The effect of this motion can therefore be eliminated by censoring the time points at the end of the imaging run.

Rationale for QC criteria: Motion censoring threshold

The correlation between functional connectivity and mean Enorm (QC-FC) was highly similar for censoring thresholds of 0.2, 0.4, and 1.0 mm (Figure 5). The mean correlation of FC with motion was close to zero (0.00001 for a motion censoring threshold

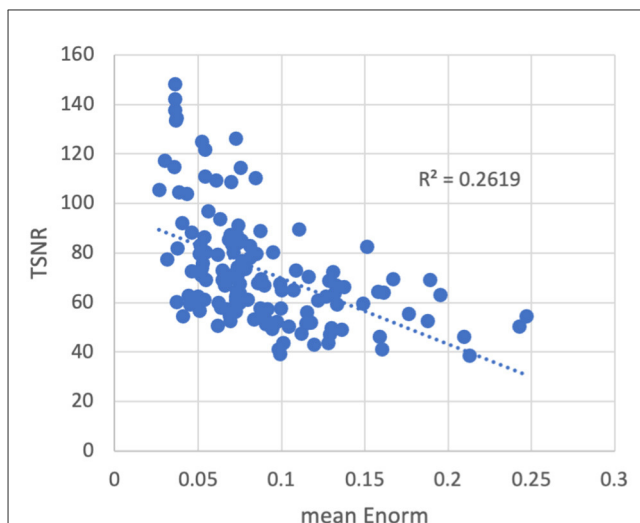


FIGURE 3

Temporal signal-to-noise ratio (TSNR) vs. the mean volume-to-volume motion as measured by the Euclidean norm (Enorm) of the temporal difference of the 6 realignment parameters. As motion increases, the TSNR decreases. Note that subjects with higher motion have lower TSNR, but the converse is not necessarily true—subjects with low motion can also have low TSNR, possibly due to other non-motion sources of noise.

of 0.2 mm, 0.001 for censoring threshold 0.4 mm, and 0.004 for a censoring threshold of 1.0 mm). The histogram showed slightly wider tails, indicating some connections with greater correlation with motion, at a censoring threshold of 1.0 mm compared to 0.4 or 0.2 mm. The QC-FC was slightly increased when no bandpass filtering was performed. There was very little distance dependence of the QC-FC. At a motion censoring threshold of 0.2 mm, the correlation between QC-FC and distance was -0.004 (95% confidence interval: -0.012 to 0.004). At a motion censoring threshold of 0.4 mm the distance dependence correlation was -0.0009 (-0.009 , 0.007), and at a motion censoring threshold of 1.0 mm the correlation was 0.005 (-0.003 , 0.013).

There was very little difference in the group functional connectivity matrices using censoring thresholds of 0.2, 0.4, or 1.0 mm (Figure 6). The similarity of each subject's functional connectivity to group mean functional connectivity was nearly the same whether the group functional connectivity matrix was formed using 0.2 vs. 0.4 mm censoring thresholds ($R^2 = 0.999$) (Figure 7). Therefore, it does not matter which motion threshold was used as the group functional connectivity for comparison in computing the similarity.

With a censoring threshold of 0.2 mm, the similarity was strongly dependent on the mean motion with lower similarity for subjects with higher motion ($R^2 = 0.27$) (Figure 8A). However, at a motion censoring threshold of 0.4 mm, the similarity was only weakly related to subject motion ($R^2 = 0.04$) (Figure 8C). The similarity of functional connectivity to the group mean was greater for a censoring threshold of 0.4 mm compared to 0.2 mm and this difference was greater in high-motion subjects (Figure 9A). That is, high-motion subjects had a higher similarity of their functional connectivity matrices to the group average using a 0.4 mm threshold compared to a more stringent 0.2 mm. This suggests

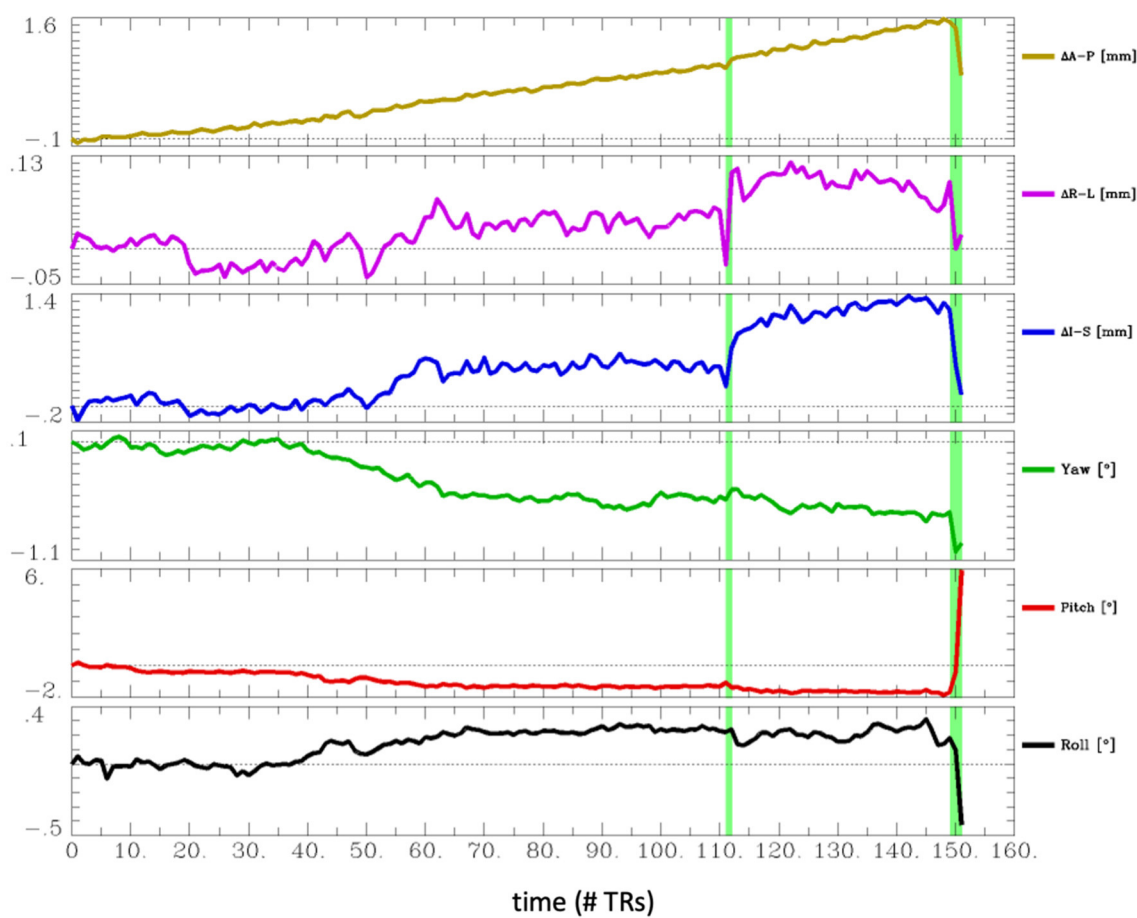


FIGURE 4
Estimated head motion realignment parameters for subject sub-102, which had the largest maximum volume-to-volume motion of 6.5 mm. However, this motion occurred at the end of the run, so the effects of this motion can be eliminated by censoring the last few time points.

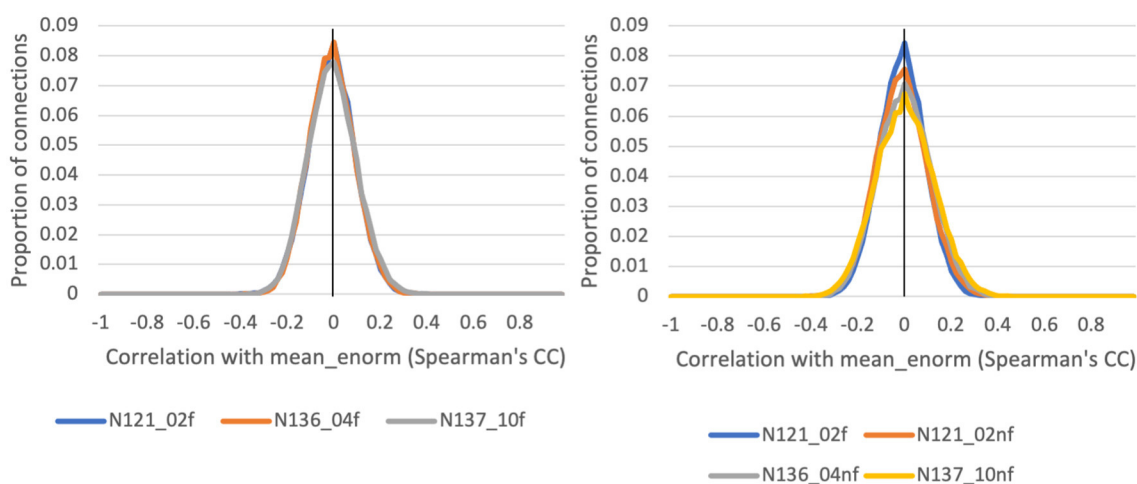
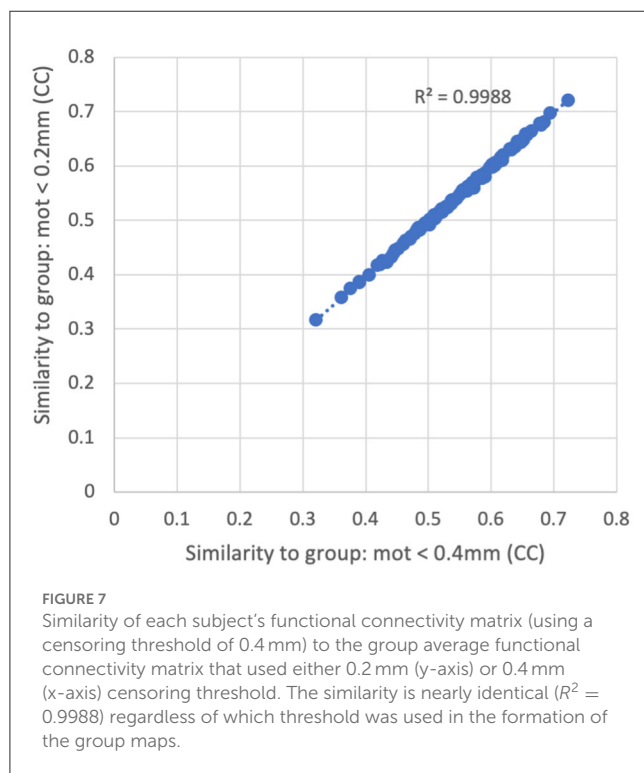
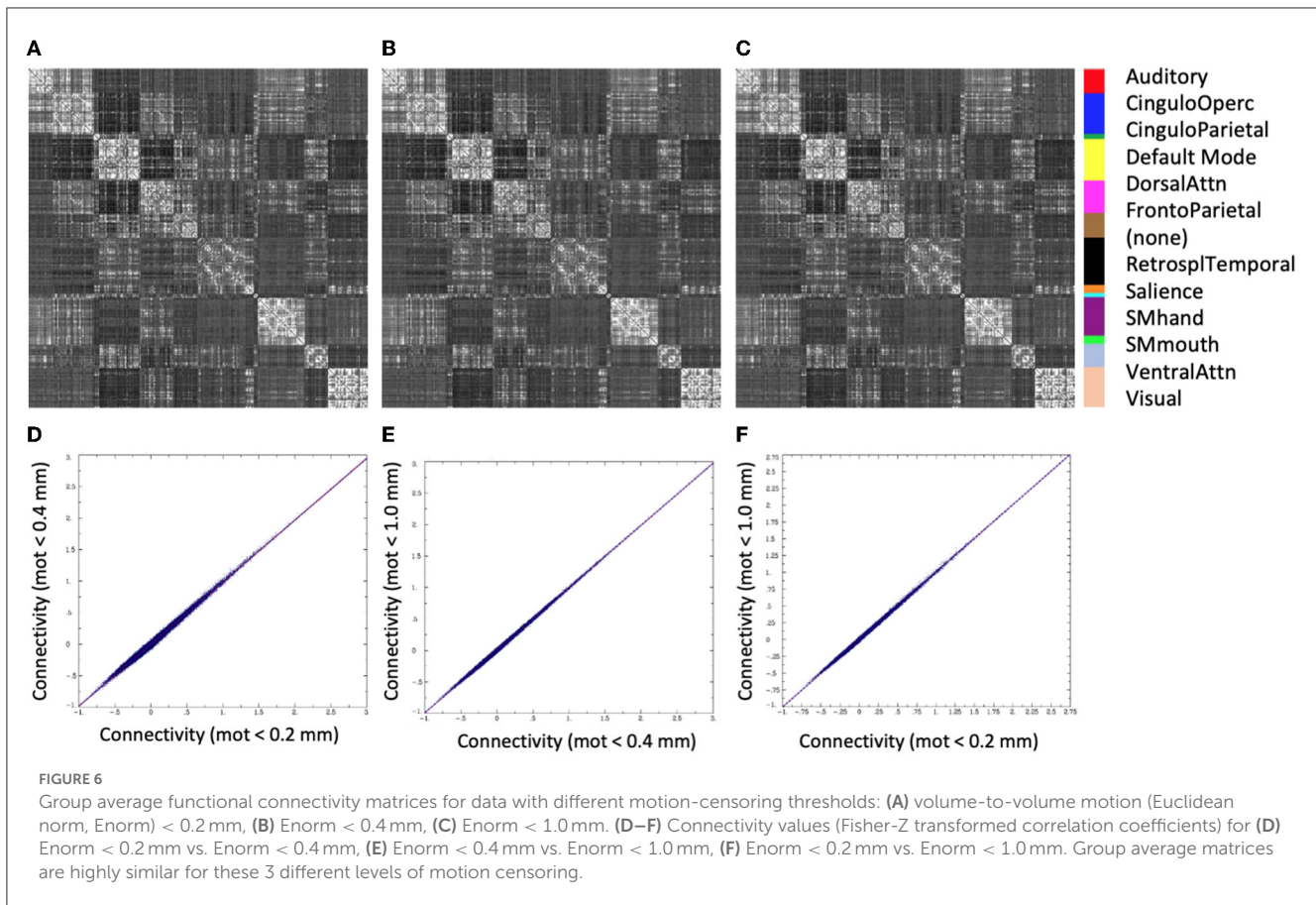


FIGURE 5
Histograms of the correlation between a quality control (QC) criterion—the mean Enorm—and the functional connectivity (FC): QC-FC, for 3 different motion censoring thresholds (02 = 0.2 mm, 04 = 0.4 mm, 10 = 1.0 mm) with (f) and without (nf) temporal bandpass filtering (0.01–0.1 Hz).

that the decreased similarity in high-motion subjects at a 0.2 mm censoring threshold is due to the reduced degrees of freedom from aggressive time point censoring rather than corruption

of the functional connectivity due to motion. Similarity was further increased, particularly in high-motion subjects, by using a motion-censoring threshold of 1.0 mm (Figure 9B). However,



this threshold is much higher than is currently used in the field, and combined with the slightly higher correlation with motion (QC-FC) at a 1.0 mm censoring threshold, we decided to use a 0.4 mm censoring threshold as the cutoff.

Figure 10 shows the similarity vs. degrees of freedom for a censoring threshold of 0.2 and 0.4 mm. Similarity is reduced for lower degrees of freedom. Moreover, there is no clear cutoff for the similarity at low degrees of freedom. The similarity appears to be roughly linearly related to the degrees of freedom for low degrees of freedom (<50), plateauing at higher degrees of freedom. We decided to use a cutoff of 15 degrees of freedom to reduce the influence of severe motion while still retaining enough subjects in the group analysis.

Similarity of functional connectivity to the group mean was also increased by eliminating the band-pass filtering step (see Figures 8, 9). One example of this is shown in Figure 11 for a subject (sub-507) that had only 7 degrees of freedom left after bandpass filtering and motion censoring with a threshold of 0.4 mm. This connectivity matrix appears quite noisy (Figure 11A). At a motion censoring threshold of 0.2 mm and no bandpass filtering, the functional connectivity matrix is more similar to the group average functional connectivity (Figure 11B). The connectivity matrix for this subject at a motion censoring threshold of 0.4 mm is very similar to a threshold of 0.2 mm when no bandpass filtering is applied

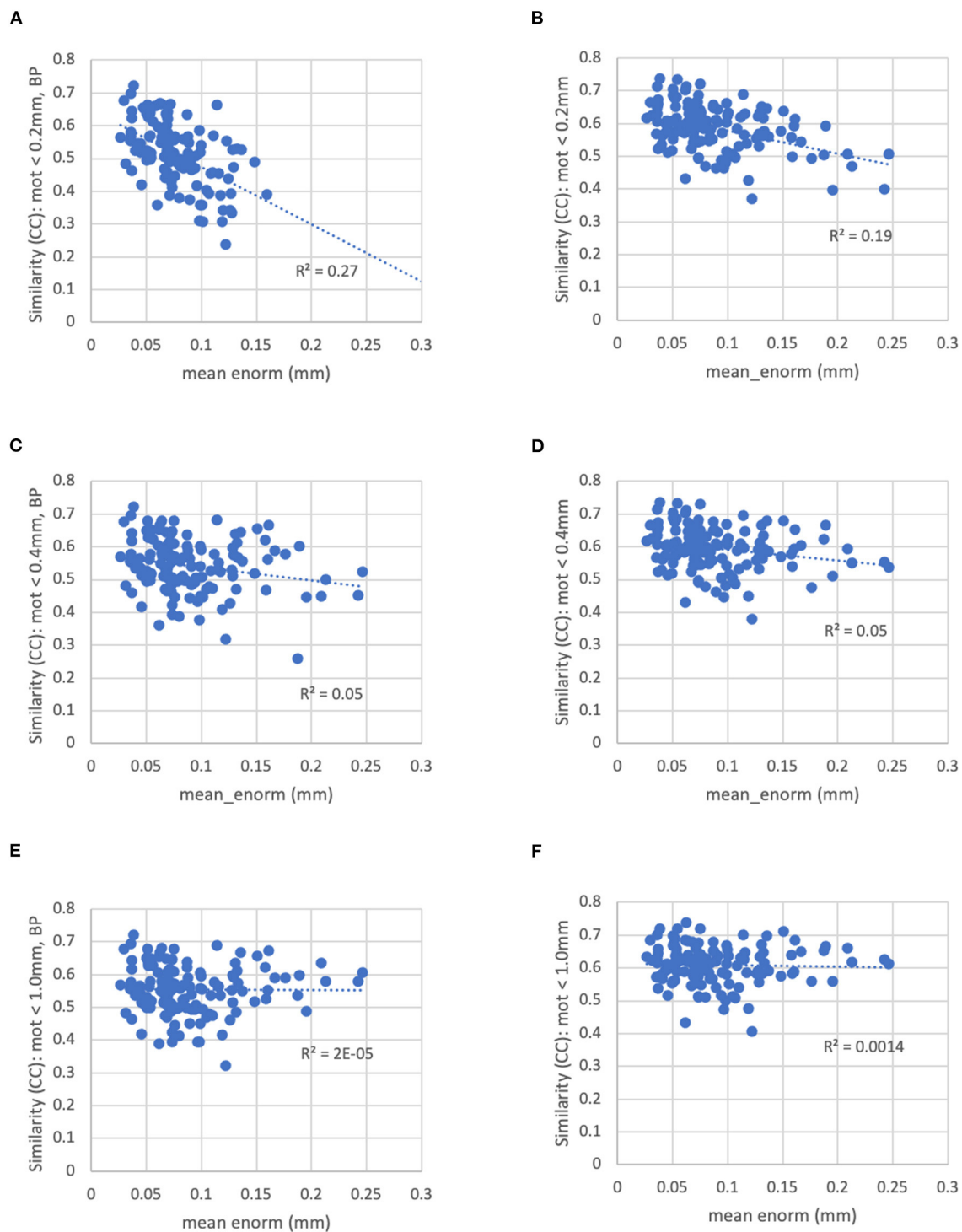


FIGURE 8

The similarity between each subject's functional connectivity matrix and the group-average functional connectivity matrix for different motion censoring thresholds (0.2, 0.4, 1.0 mm) with and without bandpass filtering (BP). BP, bandpass filtering (0.01–0.1 Hz), no BP, no bandpass filtering. **(A)** At a motion censoring threshold of 0.2 mm with bandpass filtering, subjects with higher motion (mean Enorm) show reduced similarity ($R^2 = 0.27$). **(B)** Without bandpass filtering, similarity is increased, but subjects with higher motion still show lower similarity ($R^2 = 0.19$). **(C)** At a motion censoring threshold of 0.4 mm with bandpass filtering, similarity to the group mean connectivity is only weakly correlated with motion ($R^2 = 0.05$). **(D)** Without bandpass filtering, there is again only a weak correlation with motion ($R^2 = 0.05$). **(E)** At a motion censoring threshold of 1.0 mm and bandpass filtering, there is very little correlation between the similarity and motion ($R^2 = 0.00002$). **(F)** Without bandpass filtering at a motion threshold of 1.0 mm, there is very little correlation with motion across subjects ($R^2 = 0.0014$).

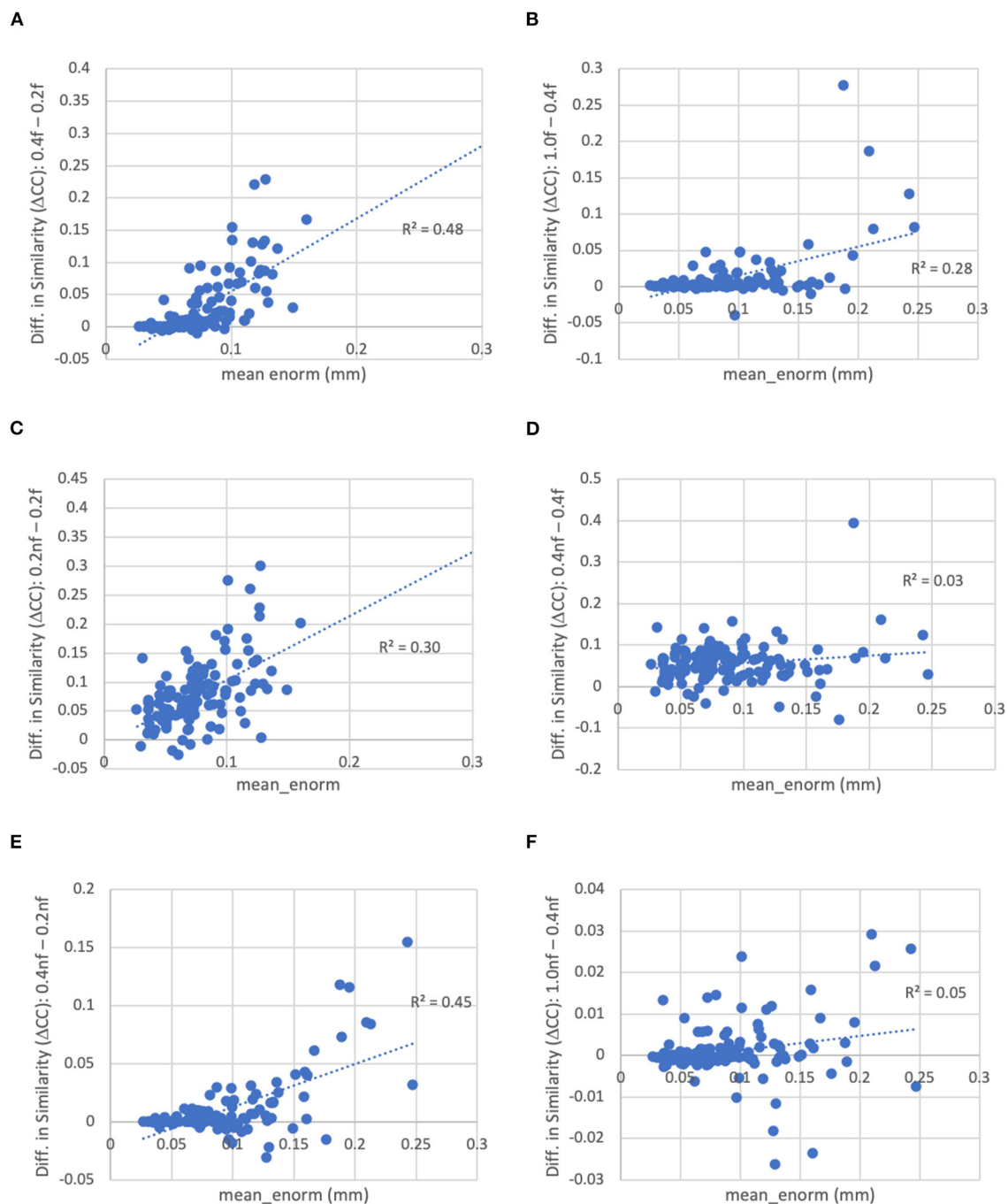


FIGURE 9

Difference in the similarity of each subject's functional connectivity matrix to the group mean for different levels of motion censoring, with (f) and without (nf) bandpass filtering. (A) With bandpass filtering, similarity is increased for connectivity matrices computed at a motion threshold of 0.4 vs. 0.2 mm, particularly in subjects with high motion. (B) Similarly with bandpass filtering, similarity is increased for a motion censoring threshold of 1.0 mm compared to 0.4 mm, particularly for high-motion subjects. (C) At a motion-censoring threshold of 0.2 mm, not performing bandpass filtering increases the similarity compared to performing bandpass filtering, particularly in high-motion subjects. (D) At a motion-censoring threshold of 0.4 mm, similarity to the group-mean is increased for most subjects without vs. with bandpass filtering, but less dependent on the mean level of motion. (E) Without bandpass filtering, a motion censoring threshold of 0.4 mm has greater similarity than a threshold of 0.2 mm, particularly for high-motion subjects. (F) Without bandpass filtering, using a motion censoring threshold of 1.0 mm compared to 0.4 mm can result in either increases or decreases in similarity to the group mean, with little correlation to mean motion.

(Figure 11C). Increase in similarity when eliminating the bandpass filtering step was observed even in low-motion subjects (Figure 12). Subject sub-501 had a mean Enorm of 0.03 mm with no time points censored at a threshold of 0.4 mm. Thirty-two degrees of freedom

were left with bandpass filtering, and 119 degrees of freedom were left without bandpass filtering. The pattern of within-network and between-network connectivity was noisier and less like the group average maps when bandpass filtering was applied. Figures 12C, D

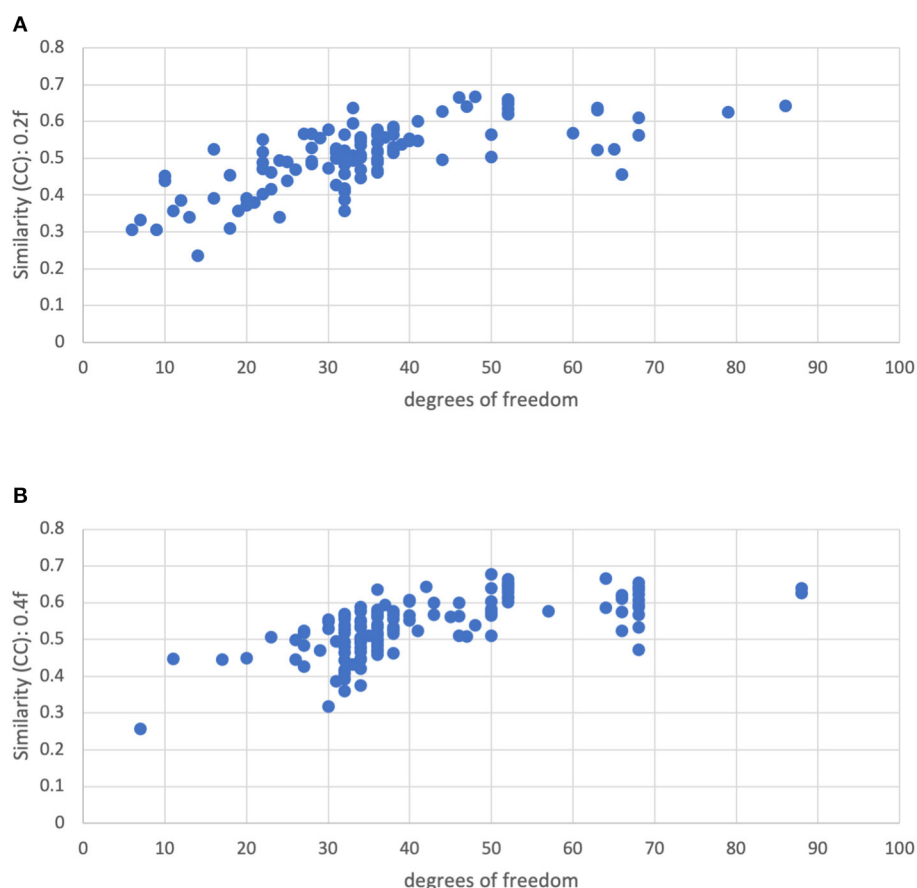


FIGURE 10

Similarity of each subject's functional connectivity matrix to the group mean for a motion censoring threshold of (Top) 0.2 mm and (Bottom) 0.4 mm, vs. the degrees of freedom left after motion censoring, bandpass filtering, and nuisance regression. The similarity appears to be roughly linearly related to the degrees of freedom for low degrees of freedom (<50), plateauing at higher degrees of freedom.

shows the connectivity matrix from a low-motion subject (sub-606) that had longer time series (720 time points), with no time points censored at a threshold of 0.4 mm, 306 degrees of freedom left after bandpass filtering and 699 degrees of freedom without bandpass filtering. Functional connectivity matrices are highly similar with and without bandpass filtering since both have high degrees of freedom.

The similarity was further improved by relaxing the motion censoring from 0.2 to 0.4 mm (Figure 9E). That is, the increase in similarity for a motion censoring threshold of 0.4 vs. 0.2 mm, both without bandpass filtering, was greater in subjects with higher *mean_enorm*, again likely due to the greater degrees of freedom with a more relaxed censoring threshold. Without bandpass filtering, the similarity was slightly correlated with *mean_enorm* at a censoring threshold of 0.2 mm ($R^2 = 0.19$, Figure 8B), but only weakly correlated with subject motion at a censoring threshold of 0.4 mm ($R^2 = 0.05$, Figure 8D). The improvements in similarity with vs. without bandpass filtering was correlated with the *mean_enorm* at a censoring threshold of 0.2 mm ($R^2 = 0.30$, Figure 9C) but not 0.4 mm ($R^2 = 0.03$, Figure 9D). These results all suggest that the similarity is improved by not applying bandpass filtering and by using a less stringent censoring threshold (e.g., 0.4 mm) due to the increased degrees of freedom. Without

bandpass filtering, using a motion censoring threshold of 1.0 mm compared to 0.4 mm resulted in either increases or decreases in similarity to the group mean for different subjects, with little correlation to mean motion (Figure 9F). Similar results were obtained when the similarity was computed using the Euclidean distance between each subject's functional connectivity matrix and the group mean rather than the Pearson's correlation (see [Supplementary material](#)).

Similar results were obtained with and without global signal regression (GSR). The similarity to the group mean functional connectivity was slightly higher with GSR, with a mean similarity (Pearson's correlation) of 0.54 with GSR compared to 0.53 without GSR ($p < 1e-12$) (see [Supplementary material](#)).

Discussion

Several datasets were identified by the quality control procedures as having deviations from expected parameters or other issues. Whether a subject should be excluded or not from further group analyses depends on the particular issue, whether this issue can be addressed, and the goals of the study. For example, excluding subjects with abnormal brain anatomy (e.g., enlarged ventricles)

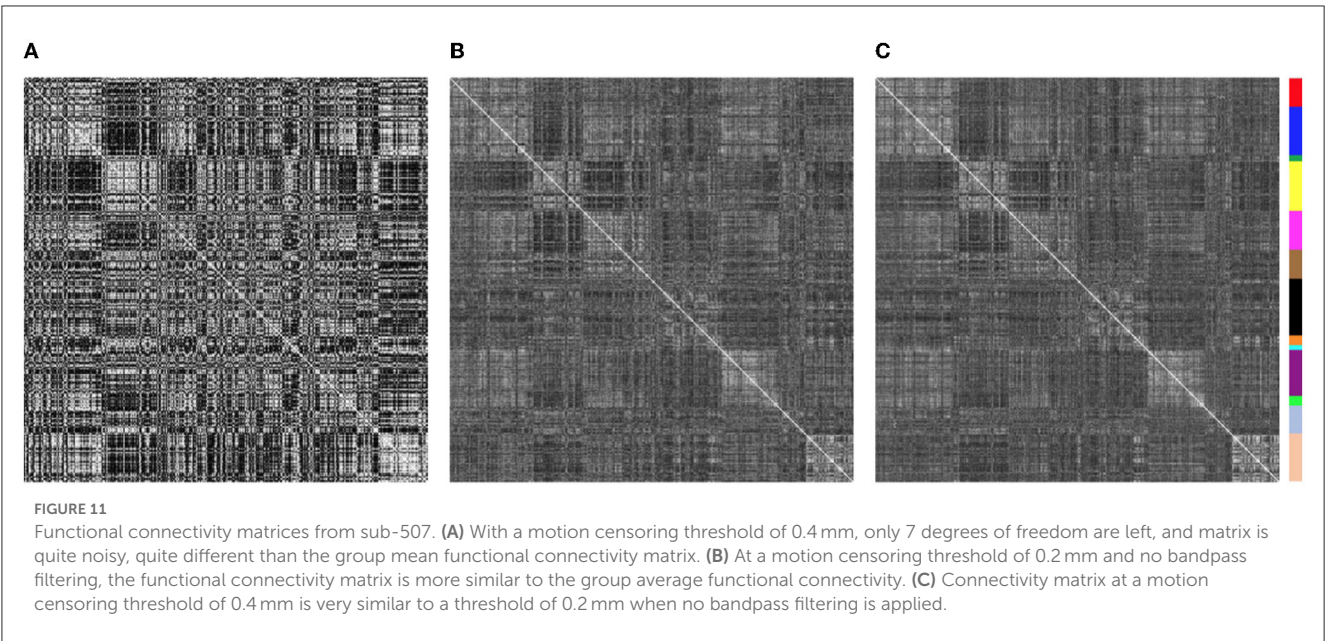


TABLE 1 QC criteria summary table.

Resting state fMRI QC criteria: Exclude (or re-examine) a subject if:
(A) Fewer than 15 degrees-of-freedom are left after motion censoring, nuisance regression, and band-pass filtering
(B) Fewer than 4 min (240 s) of data remain after motion censoring
(C) Maximum Enorm (volume-to-volume motion) > 3 mm
(D) The data are left-right flipped and the correct orientation cannot be determined
(E) Temporal signal-to-noise and/or FCS indicate the presence of an RF coil artifact (e.g., spiking)
(F) Part of the cortex is out of the field of view (qualitative)
(G) There are large abnormalities in the anatomy (qualitative)
(H) There are significant mis-alignments in the data to template space that cannot be fixed with different processing choices (qualitative)

may be advisable in studies attempting to characterize typical functional connectivity, but not in studies where such deviations are more common or of interest.

Data that had a different spatial resolution from others at that site can still be processed since all of the data are aligned and re-interpolated to a common resolution in template space, and the current study is already combining data from multiple sites which had acquired data at different spatial resolution. The data from site 500 with the higher spatial resolution (matrix size of 128 voxels vs. 80 voxels) did have greater smoothness, but the impact of this is reduced by smoothing all of the data to a similar final smoothness.

The echo-planar images from 2 subjects were flipped in the I-S direction. This may have resulted from either an error in the conversion of the DICOM files to NIFTI format, or in erroneously setting the subject position in the scanner as supine-feet-first rather than supine-head-first. This flip can in principle be easily corrected, but the process is a bit more complex since the data were acquired

with an oblique orientation. In addition, one needs to check whether the left-right orientation is also flipped. This could be done by comparing the alignment of the original and flipped versions of the EPI to the T1. Flips in the left-right orientation were identified in 2 additional subjects. It is unclear whether the error is in the EPI or the T1, but may be determined by examining the original DICOM files. These four subjects were designated as “uncertain”—if the correct left-right orientation can be determined, then they can be included; if the correct orientation cannot be determined then they should be excluded.

A motion censoring threshold of 0.2 mm is commonly used in the field. However, the findings here suggest that this threshold is too stringent for the current study, likely due to the reduced degrees of freedom with aggressive censoring. The similarity of each subject’s functional connectivity to the group mean is increased using a threshold of 0.4 mm and this similarity is no longer correlated with the mean motion, which was the case for the more stringent thresholding of 0.2 mm. Relaxing the threshold to 0.4 mm did not increase the correlation of the functional connectivity with motion (QC-FC). Similarly, there was no observable distance dependence of QC-FC at all three motion censoring thresholds evaluated.

Bandpass filtering between 0.01 and 0.1 Hz (or in some studies 0.008–0.08 Hz) is commonly performed in the field. The rationale for this processing step is that the fluctuations of interest typically occur at very low temporal frequencies (<0.1 Hz) (Biswal et al., 1995; Cordes et al., 2001), while non-neuronal fluctuations such as cardiac and respiratory fluctuations occur at much higher frequencies. However, with the typical acquisition rates (repetition times, TR), this physiological noise is aliased to lower frequencies and is not necessarily reduced by the bandpass filtering. Furthermore, bandpass filtering significantly reduces the degrees of freedom, which can affect the quality of the functional connectivity estimates (e.g., see Figure 11). The similarity of the functional connectivity to the group mean increases for nearly all subjects when no bandpass filtering is performed (this is

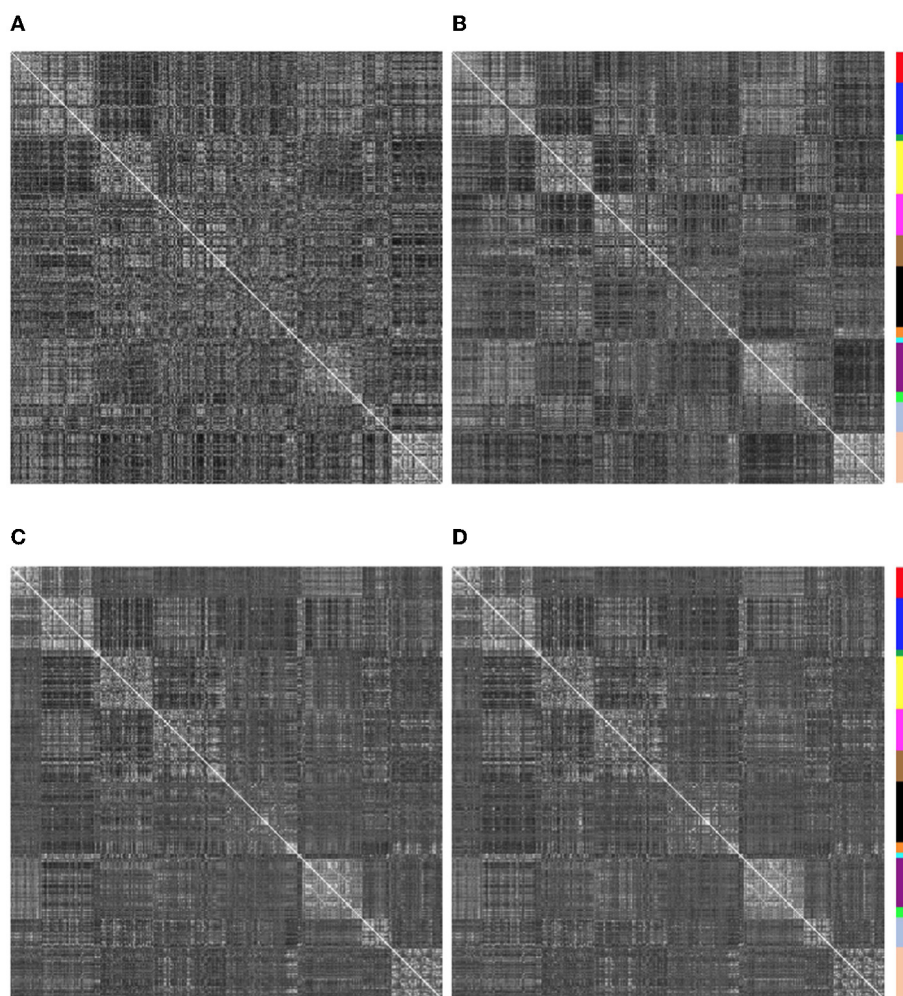


FIGURE 12

Functional connectivity matrices for 2 low-motion subjects with and without bandpass filtering (0.01–0.1 Hz). Degrees of freedom (dof) after motion censoring, nuisance regression, and with/without bandpass filtering are shown in the title. **(A, B)** Sub-501, mean Enorm = 0.03, no time points censored and 140 time points left at a censoring threshold of 0.4 mm. **(A)** With bandpass filtering, 32 degrees of freedom are left. **(B)** Without bandpass filtering, 119 degrees of freedom are left. Note that the pattern of within-network and between-network connectivity is noisier and less like the group average maps when bandpass filtering is applied. **(C, D)** Sub-606, mean Enorm = 0.03 mm, no time points censored and 720 time points left at a censoring threshold of 0.4 mm. **(C)** With bandpass filtering, 306 degrees of freedom are left. **(D)** Without bandpass filtering, 699 degrees of freedom are left. Functional connectivity matrices are highly similar with and without bandpass filtering since both have high degrees of freedom.

the case regardless of which group connectivity matrix is used for comparison—with vs. without bandpass filtering). When stringent (0.2 mm) motion censoring is applied, the similarity to the group mean is much greater without bandpass filtering compared to with bandpass filtering, particularly in higher motion subjects. This is likely due to the very low degrees of freedom in high motion subjects with both a stringent motion censoring threshold and bandpass filtering. The degrees of freedom are higher without bandpass filtering, which is likely the reason for an increase in similarity (compared to with bandpass filtering) in the higher motion subjects. At a more relaxed (0.4 mm) motion censoring threshold, the similarity does not depend on the mean motion, but is increased (by varying amounts) for nearly all subjects.

The similarity of a subject's functional connectivity to the group mean is a useful way to identify outliers and to

determine appropriate processing steps and quality control criteria (e.g., bandpass filtering, motion censoring threshold). A useful qualitative QC step is to visualize the functional connectivity maps from key seed regions (e.g., seed regions from the posterior cingulate to identify the default mode network) and see if they match the expected patterns. While not performed in the current study, quantitative metrics could be computed to measure how well the patterns of these seed-based connectivity maps match the expected pattern. An extension of this approach, in order to measure connectivity for multiple regions throughout the brain is to compute a connectivity matrix from a systematic brain-wide parcellation of the brain and examine the similarity of each subject's connectivity matrix to the group mean connectivity matrix. However, it is important to keep in mind that the goal in many functional connectivity studies is to determine the association of individual differences in functional connectivity with

some other variable. That is, we want individual differences in functional connectivity, but not those that are due to differences in subject motion. For that reason, we used the correlation of the similarity with subject motion as a guide to determine the appropriate QC criteria (motion censoring threshold), rather than using the similarity as a QC cutoff. In addition, this measure of similarity may not capture all artifacts, such as systematic errors across the entire sample.

The benefits from a motion threshold of 0.4 mm compared to 0.2 mm found here does not necessarily generalize to all other studies, in particular those acquiring much larger number of time points. The low similarity observed in many subjects in this study is due to the very low degrees of freedom remaining when more aggressive censoring is applied, particularly in combination with bandpass filtering. In studies like HCP and ABCD, where the TR is lower and many more time points have been acquired, there may be sufficient degrees of freedom left for robust estimation of functional connectivity even with more aggressive motion censoring.

The inclusion of global signal regression resulted in a statistically significant, although small, increase in the similarity of each subject's functional connectivity matrix to the group mean. This could reflect improved denoising from GSR. However, because of the lack of ground truth in resting-state functional connectivity, one should be cautious about using only QC criteria to guide processing choices. If any of the nuisance regressors (global signal, CSF signal, or white matter signal) contain effects of interest then regressing them could distort functional connectivity estimates despite improving QC metrics.

Another commonly used QC criteria is to exclude participants with large or “gross” motion, that is, if any frame-to-frame displacement exceeds a predefined threshold, such as 0.55 mm (Satterthwaite et al., 2012, 2013) or 5 mm (Parkes et al., 2018). The motivation behind this exclusion criterion is that larger motion is more likely to be associated with B0-field changes, spin-history effects, and RF coil sensitivity effects. However, if such large motion occurs relatively infrequently (e.g., only a few times during an imaging run), a recent study has shown ways to reduce the effects of this large motion (Birn et al., 2022). For this reason, the maximum motion was not used as a strict exclusion criterion in the current QC study, but simply to flag potential subjects whose functional connectivity maps should be more closely examined for potential artifacts.

Another common problem is the alignment of the EPI data to template space. Since the alignment of the T1 weighted structural images in template space was highly similar across subjects, the errors in the EPI alignment likely result from challenges in aligning the EPI to the T1. Errors in the EPI-to-template alignment were easy to identify using qualitative measures (visualization of the data), but we were not able to find any quantitative metrics that could accurately capture these errors. Misalignments between the EPI and T1 could potentially be reduced by adjusting the EPI-to-T1 alignment cost function or adjusting the parameters of the brain extraction. For example, removal of non-brain tissue (“brain extraction” or “skull-stripping”) that is too aggressive can cause clipped regions of the T1 to be stretched to fit the boundaries and gyri of the template brain. This is not often as visible on the aligned T1s (since the borders of the brain match), but can cause EPI data that is well-aligned to the T1 to be pushed outside the

template brain. The subject identified as having a misalignment was designated as “uncertain” since modified processing may result in a better alignment. Whether this subject should be excluded or included depends on the effort an investigator is willing to expend to find the processing options that result in an accurate alignment.

While the current study did not include B0-field maps, studies that do include such measures could use both qualitative and quantitative QC metrics to look at the effectiveness of B0-field distortion correction. For example, the EPI and T1 could be compared before and after correction to verify that the distortion correction was applied in the correct orientation (as determined by the phase encoding direction and polarity) and by the correct amount (as determined by the echo spacing). A Dice coefficient between the EPI and T1 could quantify this QC measure.

Qualitative measures, such as visualizing the data at different points during the processing pipeline, are an indispensable tool for quality control. One reason for this is the myriad number of ways that the processing can go awry. This quality control step can be quite time consuming, and therefore the challenge, particularly for large studies, is making this process as efficient as possible. One way to do this is to concatenate one image (e.g., T1, EPI, or connectivity map in template space) from each subject, and then scroll through the subjects manually or in a movie format. This procedure was quite useful in identifying subjects where the alignment of the EPI to template space was not ideal. These errors in alignment were not captured very well by the Dice coefficient between the EPI and T1-weighted image. This may be because the Dice coefficient between the EPI and T1 is also reduced by B0-field associated signal dropout in the orbitofrontal and temporal lobes, which vary across subjects depending on the shape of the subject's head, the angle of the head to the direction of the magnetic field, and the obliquity of the slice prescription. This signal dropout results in a lower Dice coefficient even with an accurate alignment between the EPI and T1-weighted image.

Many of the measures discussed above are provided with the QC output from the AFNI tool *afni_proc.py*. This QC output includes an alternative way to visualize the alignment of the EPI-to-T1 and T1-to-template—as outlines of the sulci and gray/white matter boundaries on top of either the EPI or the aligned T1. Since *afni_proc.py* was designed to output QC from individual subject data, it does not provide a movie of the alignment across subjects. However, such a movie could easily be generated by extracting one volume (of the EPI, T1, or connectivity map in template space) from each subject and concatenating the datasets. Alternatively, the image snapshots provided by *afni_proc*'s QC could be concatenated into a movie. Such movies can be particularly useful in identifying outliers in the alignment in a large group of subjects.

Conclusions

A number of quality control procedures and criteria are recommended for the analysis of resting-state functional MRI data. First, it is important to visualize the data at multiple points in the processing pipeline. The accuracy of alignment to template space can be evaluated by concatenating one brain volume from

each subject, and then scrolling through the subjects manually or in a movie format. Similarly, outliers in functional connectivity can be determined by concatenating functional connectivity maps from key seed regions in the brain that are known to be part of robust functional networks consistently observed across different subjects—specifically the posterior cingulate to identify the default mode network, primary motor cortex to identify the motor network, primary visual cortex to identify the visual network, and primary auditory cortex to identify the auditory network. Useful quantitative measures include the temporal signal-to-noise ratio, the degrees of freedom remaining after motion censoring and nuisance regression, and the total duration data remaining after motion censoring. While band-pass filtering of the data is currently the standard in the field, future studies may want to re-evaluate the use of this processing step particularly in studies that acquire limited amount of data. Finally, the quality control thresholds used should be examined for each study and may need to be adjusted based on the total amount of acquired data. For example, the QC cutoff of 4 min of good data and 15 degrees of freedom was based on the duration of the runs that were part of the study. Ideally one would want as much data as possible for the best reliability, but this needs to be balanced with the amount of data available and the amount of denoising desired. It is essentially a trade-off between including in the group analysis fewer subjects with “cleaner” data (fewer artifacts) or more subjects with (potentially) noisier data. The balance of this trade-off depends on the levels of motion and other artifacts and the success of noise reduction approaches.

Data availability statement

The raw data supporting the conclusions of this article are available at <https://osf.io/qaesm/> (DOI 10.17605/OSF.IO/QAESM).

References

- Birn, R. M., Dean, D. C. III, Wooten, W., Planalp, E. M., Kecsckemeti, S., Alexander, A. L., et al. (2022). Reduction of motion artifacts in functional connectivity resulting from infrequent large motion. *Brain Connect* 12, 740–753. doi: 10.1089/brain.2021.0133
- Birn, R. M., Molloy, E. K., Patriat, R., Parker, T., Meier, T. B., Kirk, G. R., et al. (2013). The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *Neuroimage* 83, 550–558. doi: 10.1016/j.neuroimage.2013.05.099
- Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409
- Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Ciric, R., Rosen, A. F. G., Erus, G., Cieslak, M., Adebimpe, A., Cook, P. A., et al. (2018). Mitigating head motion artifact in functional connectivity MRI. *Nat. Protoc.* 13, 2801–2826. doi: 10.1038/s41596-018-0065-y
- Cordes, D., Haughton, V. M., Arfanakis, K., Carew, J. D., Turski, P. A., Moritz, C. H., et al. (2001). Frequencies contributing to functional connectivity in the cerebral cortex in “resting-state” data. *AJNR Am. J. Neuroradiol.* 22, 1326–1333. Available online at: <https://pubmed.ncbi.nlm.nih.gov/11498421/>
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Di Martino, A., Yan, G.-C., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., et al. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* 18, 1664–1671. doi: 10.1038/nn.4135
- Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinsty, R. C., Collins, D. L., et al. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* 54, 313–327. doi: 10.1016/j.neuroimage.2010.07.033
- Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., et al. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Hum. Brain Mapp.* 29, 958–972. doi: 10.1002/hbm.20440
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312
- Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., Petersen, S. E., et al. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* 26, 288–303. doi: 10.1093/cercor/bhu239

Author contributions

RB conceived of the ideas, performed the analyses, and wrote the manuscript.

Funding

This research was supported in part by NIH grant R01 MH128371.

Conflict of interest

RB is a consultant for NOUS, Inc.

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnimg.2023.1072927/full#supplementary-material>

- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *Elife* 10, e71774. doi: 10.7554/eLife.71774
- Parkes, L., Fulcher, B., Yucel, M., and Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *Neuroimage* 171, 415–436. doi: 10.1016/j.neuroimage.2017.12.073
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., et al. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* 84, 320–341. doi: 10.1016/j.neuroimage.2013.08.048
- Power, J. D., Schlaggar, B. L., and Petersen, S. E. (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage* 105, 536–551. doi: 10.1016/j.neuroimage.2014.10.044
- Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughhead, J., Calkins, M. E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage* 64, 240–256. doi: 10.1016/j.neuroimage.2012.08.052
- Satterthwaite, T. D., Wolf, D. H., Loughhead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., et al. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage* 60, 623–632. doi: 10.1016/j.neuroimage.2011.12.063
- Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908
- Van Dijk, K. R., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W., Buckner, R. L., et al. (2010). Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *J. Neurophysiol.* 103, 297–321. doi: 10.1152/jn.00783.2009
- Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* 59, 431–438. doi: 10.1016/j.neuroimage.2011.07.044
- Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424



OPEN ACCESS

EDITED BY

Paul A. Taylor,
National Institute of Mental Health (NIH),
United States

REVIEWED BY

Pradeep Reddy Raamana,
University of Pittsburgh,
United States
Elena Pozzi,
The University of Melbourne,
Australia

*CORRESPONDENCE

Alfonso Nieto-Castañón
✉ alfnie@bu.edu

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

RECEIVED 07 November 2022

ACCEPTED 01 March 2023

PUBLISHED 23 March 2023

CITATION

Morfini F, Whitfield-Gabrieli S and
Nieto-Castañón A (2023) Functional
connectivity MRI quality control procedures in
CONN.
Front. Neurosci. 17:1092125.
doi: 10.3389/fnins.2023.1092125

COPYRIGHT

© 2023 Morfini, Whitfield-Gabrieli and Nieto-Castañón. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Functional connectivity MRI quality control procedures in CONN

Francesca Morfini¹, Susan Whitfield-Gabrieli^{1,2,3} and
Alfonso Nieto-Castañón^{4,5*}

¹Department of Psychology, Northeastern University, Boston, MA, United States, ²Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States, ³Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, United States, ⁴McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, United States, ⁵Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA, United States

Quality control (QC) for functional connectivity magnetic resonance imaging (FC-MRI) is critical to ensure the validity of neuroimaging studies. Noise confounds are common in MRI data and, if not accounted for, may introduce biases in functional measures affecting the validity, replicability, and interpretation of FC-MRI study results. Although FC-MRI analysis rests on the assumption of adequate data processing, QC is underutilized and not systematically reported. Here, we describe a quality control pipeline for the visual and automated evaluation of MRI data implemented as part of the CONN toolbox. We analyzed publicly available resting state MRI data ($N=139$ from 7 MRI sites) from the FMRI Open QC Project. Preprocessing steps included realignment, unwarp, normalization, segmentation, outlier identification, and smoothing. Data denoising was performed based on the combination of scrubbing, motion regression, and aCompCor – a principal component characterization of noise from minimally eroded masks of white matter and of cerebrospinal fluid tissues. Participant-level QC procedures included visual inspection of raw-level data and of representative images after each preprocessing step for each run, as well as the computation of automated descriptive QC measures such as average framewise displacement, average global signal change, prevalence of outlier scans, MNI to anatomical and functional overlap, anatomical to functional overlap, residual BOLD timeseries variability, effective degrees of freedom, and global correlation strength. Dataset-level QC procedures included the evaluation of inter-subject variability in the distributions of edge connectivity in a 1,000-node graph (FC distribution displays), and the estimation of residual associations across participants between functional connectivity strength and potential noise indicators such as participant's head motion and prevalence of outlier scans (QC-FC analyses). QC procedures are demonstrated on the reference dataset with an emphasis on visualization, and general recommendations for best practices are discussed in the context of functional connectivity and other fMRI analysis. We hope this work contributes toward the dissemination and standardization of QC testing performance reporting among peers and in scientific journals.

KEYWORDS

fMRI, quality control, neuroimaging (anatomic), CONN toolbox, functional connectivity, resting state, preprocessing, denoising

1. Introduction

Since its inception, neuroimaging has escalated our understanding of the brain in both health and disease. Functional magnetic resonance imaging (fMRI) is among the most common neuroimaging techniques, as it allows us to approximate neural activity *in vivo* and non-invasively by measuring the blood oxygenation level-dependent (BOLD) signal. Brain functional connectivity (FC), or the temporal coupling of BOLD signals from anatomically distant regions, is widely used to probe neural functioning, neurodiversity, and their relationship with behavior during explicit or implicit (i.e., at rest) tasks. However, the BOLD signal is noisy and only marginally representative of neural activity. It is generated from complex interactions between neuronal, metabolic, cardiac, vigilance, and other physiological processes (Bianciardi et al., 2009; Liu, 2016; Liu and Falahpour, 2020) and is commonly affected by machine-related and participant-specific characteristics. In many fMRI analyses, these noise sources act as nuisance effects, increasing variability of the BOLD signal and ultimately reducing the power and replicability of fMRI analysis results. In functional connectivity analyses, their effect is considerably more damaging, as many of these noise sources are highly correlated across different areas and will bias functional connectivity estimates, acting as confounder effects and affecting the validity and interpretation of FC-MRI analysis results.

Commonly, anatomical and functional data undergo a series of transformations aimed at minimizing the effects of these well-known sources of BOLD signal variability prior to statistical analysis. Functional and anatomical data are usually first preprocessed with a set of steps addressing mainly spatial properties of the data that are a direct consequence of the specificities of the fMRI acquisition procedure. Specifically, preprocessing focuses on intra-participant coregistration, e.g., compensating for head motion across different functional scans, correcting for inter-slice temporal differences and magnetic susceptibility distortions, when appropriate, as well as inter-participant coregistration, e.g., by spatially projecting each subject's anatomy to a common reference space. However, despite these common preprocessing steps, functional timeseries after preprocessing usually still contain substantial variability associated with non-neural sources, including cardiac, respiratory, and residual subject motion effects, limiting the ability to effectively use these data for statistical analyses without additional control or correction of these factors. For these reasons, and particularly in the context of functional connectivity analyses, preprocessed functional timeseries are often usually then denoised by a combination of band-pass filtering and regression of temporal components characterizing these additional noise sources. Many effective alternatives have been suggested to achieve optimal preprocessing (Friston et al., 1996; Murphy et al., 2009; Chai et al., 2012; Hallquist et al., 2013; Power et al., 2014; Ciric et al., 2017) and denoising performance (Parkes et al., 2018; Maknojia et al., 2019; Tong et al., 2019; De Blasi et al., 2020; Golestani and Chen, 2022; for a review, see Caballero-Gaudes and Reynolds, 2017). Regardless of the specific pipelines applied, preprocessing and denoising have been shown to successfully reduce the effect of known nuisance factors.

However, the beneficial effect of preprocessing and denoising depends on the ability of each step to successfully achieve its intended goal. Quality control (QC) procedures are designed to evaluate the

quality of the data and to detect potential problems either in the original data or arising from failed or insufficient preprocessing and denoising steps. Quality control is an integral part of preparing fMRI data for statistical analyses, as without it there is no meaningful way to avoid problems in the data from affecting statistical analyses, leading to results that may fail to replicate, may be disproportionately influenced by the presence of outliers, or may be confounded by physiological or other non-neural sources of variability among participants. While data quality is an agreed-upon essential element for fMRI analysis, what constitutes “good” data and “appropriate” QC procedures are still open questions. Perhaps owing to the complexity of assessing data quality in the absence of a ground truth, QC is often underappreciated and not systematically reported. Yet, QC and QC reporting are crucial to data interpretation and needed to develop standardized guidelines (Taylor et al., 2022).

Several studies have addressed the topic of MRI data quality, whether from the perspective of quality assurance (QA) or from a QC point of view. Although interwoven, QA and QC are complementary in that QA is usually a process-oriented approach aimed at preventing issues (e.g., Friedman and Glover, 2006; Glover et al., 2012; Liu et al., 2015; for a review see Lu et al., 2019), whereas QC is output-oriented and evaluates the quality of the images resulting from said process. As such, even an optimal QA does not address the objectives of QC testing. Recent efforts from the field have resulted in the proliferation of QC tools and protocols for the evaluation of specific analytical step (Backhausen et al., 2016; Storelli et al., 2019; Benhajali et al., 2020), pipelines-specific outputs (Griffanti et al., 2017; Raamana et al., 2020; Chou et al., 2022), and raw-level data [e.g., MRIQC (Esteban et al., 2017) and pyfMRIqc (Williams and Lindner, 2020)]. Additionally, many pipelines have been developed to preprocess (e.g., fMRIPrep; Esteban et al., 2019), denoise (e.g., Tedana; DuPre et al., 2021), or generally analyze fMRI data from specific consortia [e.g., ABCD (Hagler et al., 2019), UK Biobank (Alfaro-Almagro et al., 2018), HCP (Marcus et al., 2013), Configurable Pipeline for the Analysis of Connectomes C-PAC¹ (Craddock et al., 2013; Sikka et al., 2014)]. While principally focused on data analysis, these tools also strongly support automatic and visual QC, and effectively aid the identification of issues in the data and during data analysis. These works, together with our and the other papers presented in this special issue (Taylor et al., 2022), help build a rich diversity of approaches and perspectives. Each provides unique contributions which help expand the field and build a consensus on best practices.

In this study, we describe the quality control pipeline for volume-based connectivity analysis using the CONN toolbox (Whitfield-Gabrieli and Nieto-Castanon, 2012; Nieto-Castanon, 2020). We analyzed publicly available resting-state data ($n = 139$) from the FMRI Open QC Project (Taylor et al., 2022) to demonstrate participant-level and group-level QC procedures in an integrated framework with data preprocessing and denoising. Visual and automated QC procedures were demonstrated for the assessment of raw-level, preprocessed, and denoised data. Finally, we proposed a QC workflow based on the combination of visual and automated QC measures. Ultimately, we hope this work contributes toward the dissemination and standardization of QC testing and reporting.

¹ <https://www.nitrc.org/projects/cpac/>

2. Materials

2.1. Dataset overview

We analyzed data from the FMRI Open QC Project (Taylor et al., 2022) fmri-open-qc-rest collection v1.0.0, which combined subsamples of public data-packages including ABIDE and ABIDE-II (Di Martino et al., 2013), the Functional Connectome Project (Biswal et al., 2010), and OpenNeuro (Markiewicz et al., 2021). Data was accessed as already transformed nifti and json files curated to be in BIDS format v1.6.0 (Gorgolewski et al., 2016).

The fmri-open-qc-rest collection included (f)MRI data from 139 participants acquired with 3.0T MRI scanners from seven sites. Each participant had available data corresponding to one MRI scanning session when one anatomical image and one or two echo-planar imaging (EPI) resting state functional BOLD runs were collected.

2.2. Software information

MRI data processing and statistical analyses were performed using the CONN toolbox (RRID:SCR_009550) version 22.a (Nieto-Castanon and Whitfield-Gabrieli, 2022) and SPM version 12 release 7,771 (Wellcome Department of Imaging Neuroscience, UCL, London, United Kingdom) in MATLAB R2022a (The MathWorks Inc., Natick, MA, United Kingdom).

3. Methods

Code and scripts required to replicate the analysis presented in this manuscript can be found at <https://github.com/alfnie/conn>.

3.1. Preprocessing

Functional and anatomical images were preprocessed using the default minimal preprocessing pipeline in CONN (Nieto-Castanon, 2020, 2022), represented in Figure 1 (top). This pipeline includes functional *realignment and unwarp* (Andersson et al., 2001) for intermodality coregistration of all scans to the first scan, *slice-timing correction* (STC; Henson et al., 1999) compensating for acquisition time differences among different slices, *outlier detection* (Whitfield-Gabrieli et al., 2011) identifying individual scans with suprathreshold framewise displacement (FD) and/or global signal change (GSC) values, *direct functional normalization* (Calhoun et al., 2017) projecting functional images into standard Montreal Neurological Institute 152 (MNI) reference space resampled to 2 mm isotropic voxels, and spatial *smoothing* with a 8 mm full width at half maximum Gaussian kernel. Anatomical data preprocessing comprised *direct segmentation and normalization* (Ashburner and Friston, 2005) which iteratively performed tissue *segmentation* into six tissue classes, including gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using SPM12 posterior tissue probability maps, and *normalization* to IXI-549 MNI space, resampling the output anatomical images to 1 mm isotropic voxels.

Several automated measures were extracted as run-level timeseries (i.e., as 1st-level covariates) at various stages of preprocessing,

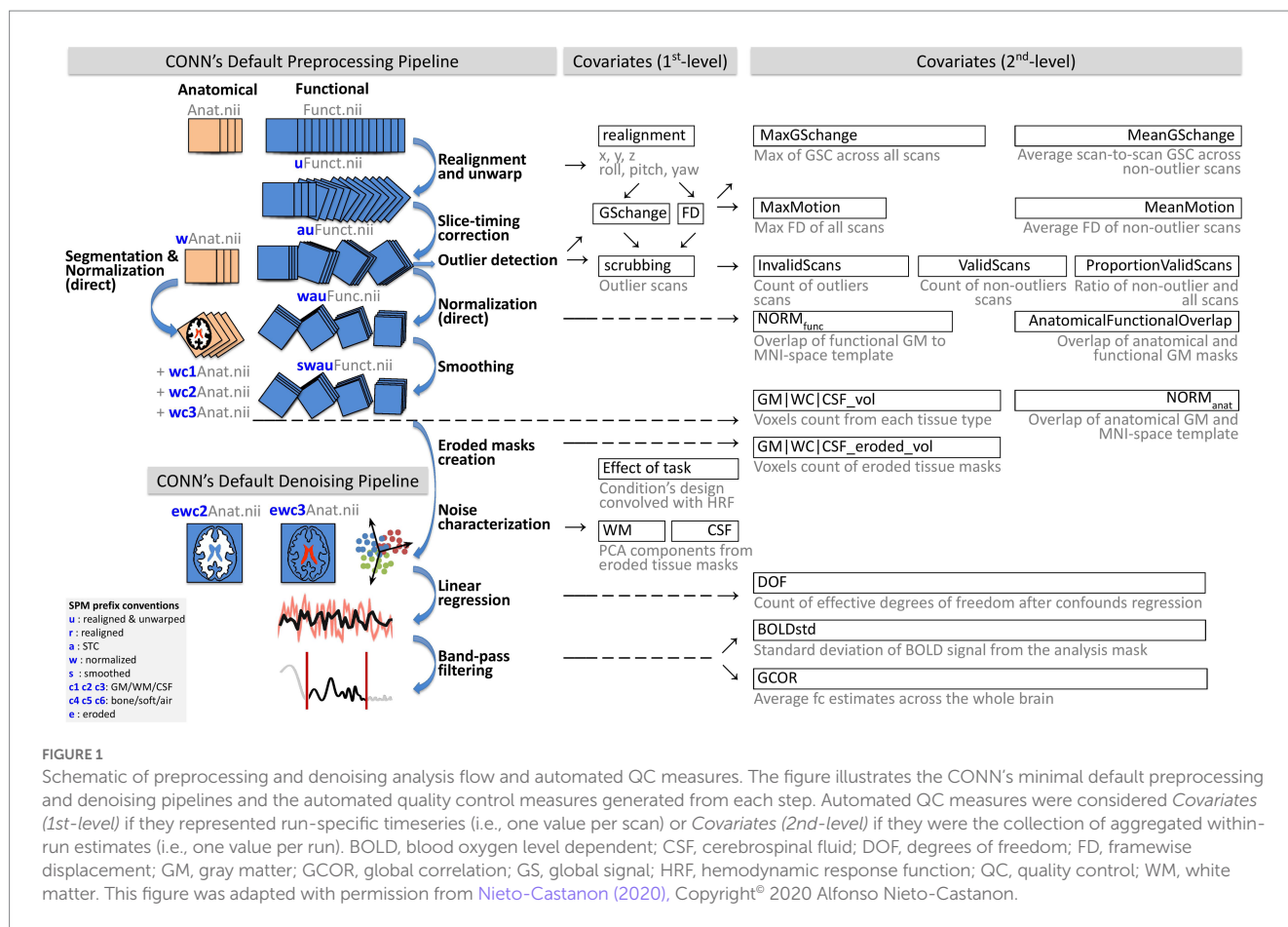
following Nieto-Castanon (2020, 2022). Table 1 (QC timeseries section) includes a summary of each of these QC timeseries definitions, and Figure 1 provides a schematic representation of all preprocessing steps and associated QC timeseries. The QC timeseries named *realignment* is estimated during the *realignment and unwarp* preprocessing step, and it represents the estimated participant in-scanner head motion. The individual parameters in this timeseries represent the degree of relative translation (three parameters, in mm units) and rotations (three parameters, in radians) of the head at each individual scan, when compared to its position at the beginning of the functional run. Following SPM12 convention, rotation parameters are defined using the real word-space point (coordinate 0,0,0) as the center of rotation. The QC timeseries named Global Signal Change (GSC) and Framewise Displacement (FD) are computed during the *outlier detection* preprocessing step. GSC timeseries are defined at each scan as the absolute value of the scan-to-scan change in global BOLD signal, using SPM global BOLD signal definition. GSC timeseries are then scaled to standard units within each run by subtracting their median value and dividing by 0.74 times their interquartile range (Whitfield-Gabrieli et al., 2011). FD timeseries are defined as the maximum change in the position of six points placed at the centers of each face in a 140 × 180 × 115 mm bounding box around the brain and undergoing the same rotations and translations as the participant's head. From these measures, outlier scans are identified as the scans with FD values above 0.5 mm and/or GSC values above 3 standard deviations (Whitfield-Gabrieli et al., 2011), with the resulting list of potential outlier scans summarized in the QC timeseries named scrubbing.

In addition to being useful on their own to characterize image and subject properties during data acquisition in the scanner, relevant statistics of these 1st-level measures are also used to define additional summary measures, as shown in Table 1 (QC summary measures section) and discussed in section 3.3.2.

3.2. Denoising

In order to minimize the presence of non-neural noise sources, including cardiac, respiratory, and residual subject motion effects in the BOLD signal, functional data were denoised with the CONN fMRI default denoising pipeline (Nieto-Castanon, 2020). This pipeline comprises three main sequential steps (Figure 1, bottom) seeking to characterize noise components in the BOLD signal (*noise components extraction*) and minimize their effect on the BOLD timeseries (*linear regression* and *temporal band-pass filtering* steps). First, participant-specific minimally eroded WM and CSF masks were generated using a one-voxel binary 3D erosion of the corresponding tissue masks derived from each subject's anatomical *segmentation*. The QC timeseries named WM and CSF (Table 1) are defined as the principal components of the BOLD signal extracted from these minimally eroded masks, following the anatomical aCompCor method (Behzadi et al., 2007), which has been shown to minimize the effect of nuisance confounds (Chai et al., 2012). Principal components from WM and CSF areas were computed after discounting motion and outlier effects (within a space orthogonal to the realignment and scrubbing QC timeseries).

Next, ordinary least squares regression removed from each voxel BOLD timeseries the effect of all identified noise components,



including 5 components from white matter (from the QC timeseries WM), 5 components from CSF (from the QC timeseries CSF), 12 estimated participant-motion parameters (6 parameters from the QC timeseries realignment and their first order temporal derivatives), participant-specific outlier scans (from the QC timeseries scrubbing), as well as the effect of session and its first order derivative convolved with the canonical hemodynamic response function (aiming to minimize the influence of transients in the first few scans of each run), and constant and linear session effects (aiming to minimize the influence of linear trends in each run). Lastly, temporal band-pass filtering (0.008–0.09 Hz) was applied to each run individually (Hallquist et al., 2013) in order to focus on slowly varying BOLD signal fluctuations.

3.3. CONN quality control pipeline

QC of raw-level, preprocessed, and denoised data was carried out following CONN quality control pipeline, building off from Nieto-Castanon (2020, 2022) and summarized in Figure 2.

3.3.1. Quality control of raw-level data

Raw-level functional runs (all slices and all scans) and anatomical images (all slices) were visually inspected using multislice interactive displays of each participant's data, as well as a combined montage of a single slice across all participants. We also inspected information from json sidcar files and header of nifti files to gather information about

image resolution and scanner acquisition parameters. The goal of this step was to familiarize ourselves with the data, identify potential sources of heterogeneity, possible incongruencies among different sites or subjects, and inspect the data for potential outliers or artifacts that may require additional consideration during preprocessing.

3.3.2. Quality control of preprocessed data

Plots of representative brain slices and automated QC measures were generated for each individual subject and functional run to visualize the outputs of preprocessing, identify potential failures of functional and anatomical preprocessing steps, or otherwise confirm that between-run spatial heterogeneity across subjects and runs had been in fact minimized as a result of these steps.

Visual QC included the assessment of the accuracy of functional normalization through the inspection of plots rendering the mean BOLD signal across all scans of the normalized functional data for each participant overlaid onto the 25% boundaries of the gray matter *a priori* probability maps from SPM's IXI-549 MNI-space template. Similarly, the accuracy of structural normalization was assessed through the inspection of plots displaying each participant's normalized anatomical images overlaid onto the same gray matter boundaries. Segmentation and anatomical to functional alignment were assessed through plots overlaying the boundaries of each participant's anatomical GM masks onto the normalized anatomical or functional data.

The presence of potential residual artifacts in functional timeseries was reviewed based on plots displaying a movie of the central axial slice (MNI $z = 0$ mm) of the functional data over time

TABLE 1 Summary of automated quality control measures.

QC timeseries (1st-level covariates)		
GSchange	The global signal change timeseries is computed as the absolute value of the scan-to-scan change in global BOLD signal, computed separately at each scan/timepoint and scaled to standard units within each run.	$0 < x < \infty$. Higher values indicate higher sudden variability in signal intensity.
FD	The framewise displacement timeseries is computed as the maximum change in the position of six control points placed at the center of a bounding box around the brain, computed separately at each scan/timepoint.	$0 < x < \infty$. Higher values indicate higher sudden displacements in head position.
Scrubbing	The scrubbing covariate contains one separate timeseries per identified outlier scan. Each of these timeseries contain a single 1-value at the identified scan, and 0-values at all other timepoints. They are computed by thresholding GSchange and FD at the desired values.	$x \in \{0,1\}$. 1 indicates a scan identified as a potential outlier
Realignment	The realignment covariate contains six timeseries, three characterizing head translations along the x/y/z directions in mm units, and three characterizing rotations around the x/y/z axes in radians.	$-\infty < x < \infty$. Higher absolute values indicate larger relative motion between a scan compared to the first scan within the same run
WM	The WM covariate contains multiple timeseries, characterizing the principal components of the BOLD signal within white matter areas, sorted by decreasing variance.	$-\infty < x < \infty$. Higher absolute values indicate larger departures from the average BOLD signal within WM
CSF	The CSF covariate contains multiple timeseries, characterizing the principal components of the BOLD signal within cerebrospinal fluid tissue areas, sorted by decreasing variance.	$-\infty < x < \infty$. Higher absolute values indicate larger departures from the average BOLD signal within CSF
QC summary measures (2nd-level covariates)		
MaxMotion	The maximum of motion is the maximum value of the FD timeserie from each run, calculated considering all original scans.	$0 < x < \infty$. Higher values indicate more extreme motion spikes.
InvalidScans	Invalid scans is the number of scans identified as outliers during outlier detection based on scan-to-scan GS and framewise displacement change.	$0 < x < \text{total number of scans}$. Higher values indicate higher presence of potential outlier scans.
ValidScans	Valid scans is the number of valid or non-outlier scans.	$0 < x < \text{total number of scans}$. Lower values indicate fewer surviving scans.
PVS	The proportion of valid scans is the ratio between non-outlier scans to all scans, representing a normalized measure of valid scans in the presence of potential differences in scanning lengths.	$0 < x < 1$. Lower values indicate higher presence of potential outlier scans.
MeanGSchange	The mean global signal change is the mean value of GSchange timeseries, calculated by aggregating GSchange across non-outlier scans only.	$-\infty < x < \infty$. Higher values indicate higher residual variability in the global signal after scrubbing
MeanMotion	The mean motion is the mean value of the FD timeseries, calculated by aggregating FD across non-outlier scans only.	$0 < x < \infty$. Higher values indicate higher residual motion after scrubbing.
$NORM_{func}$	The normalized space to functional accuracy is the Dice similarity coefficient between the IXI-549 MNI-space gray matter tissue mask thresholded at a 25% probability level and the binarized GM masks derived from the functional data and thresholded at a level that produced the same number of suprathreshold voxels as in the MNI-space mask.	$0 < x < 1$. Lower values indicate a worse normalization of functional data.
$NORM_{anat}$	The normalized space to anatomical accuracy is calculated similarly to $NORM_{func}$ but it compares the IXI-549 gray matter mask to the binarized GM mask derived from the anatomical data instead.	$0 < x < 1$. Lower values represent worse normalization of anatomical data.
AFO	The anatomical-to-functional overlap is the Dice similarity coefficient between the anatomical gray matter mask, thresholded at a 50% probability level, and the functional gray matter mask, thresholded at a level that resulted in the same number of suprathreshold voxels.	$0 < x < 1$. Lower values represent a worse inter-modality coregistration.
tissue_vol	The gray matter, white matter, or cerebrospinal fluid tissue volumes is the count of voxels with tissue-specific probability >50% from participant-specific segmented anatomical tissue ROIs.	$0 < x < \infty$. Extreme values indicate a combination of individual anatomical differences and normalization performance.
tissue_eroded_vol	The tissue eroded volume is the count of voxels in the tissue-specific ROIs resulting from anatomical segmentation after a 1-voxel erosion procedure.	$0 < x < \infty$. Extreme values indicate a combination of individual anatomical differences and normalization performance.
DOF	The effective degrees of freedom are calculated as the total number of scans minus the number of regressors involved in the denoising's linear regression step, multiplied by the fraction of the Nyquist frequency covered by denoising's band-pass frequency filter.	$-\infty < x < \text{all original scans}$. Lower values indicate potential lack of precision when estimating modeled effects in the BOLD signal.

(Continued)

TABLE 1 (Continued)

QC summary measures (2nd-level covariates)		
BOLDstd	The BOLD standard deviation is the temporal standard deviation of the BOLD signal, after grand-mean scaling to 100 across the entire brain and denoising, averaged across all runs and all voxels in the analysis mask.	$0 < x < \infty$. High values may indicate the presence of potential noise, while values close to 0 may indicate lack of retained signal.
GCOR	The mean global correlation (Saad et al., 2013) is the average of Pearson's r correlation coefficients between the denoised BOLD timeseries of all pairs of voxels within the analysis mask.	$-\infty < x < \infty$. High absolute values may indicate the presence of residual noise sources in the BOLD signal.
QC-FC %	Quality Control to Functional Connectivity distributions (Ciric et al., 2017) represent the observed distribution of correlations across participants between individual QC measures and functional connectivity strength (edges in a fixed graph of 1,000 random voxels within the MNI-space gray matter template mask). QC-FC % match level represents the distance between these observed distributions and those that could be expected by chance, as computed using permutation analyses.	$0\% < x < 100\%$. Values above 95% indicate negligible modulations associated with nuisance factors in the correlation structure of the BOLD signal.

All quality control measures are automatically calculated by CONN (v22.a) during data preprocessing and denoising, but all could also be derived post-hoc from data fully or partially processed by other software.

(i.e., over all scans). This movie was rendered above the timeseries traces of (i) the GSC QC timeseries representing scan-to-scan changes in the global BOLD signal, (ii) the FD QC timeseries, characterizing subject motion, and (iii) the outlier QC timeseries, characterizing scans identified as potential outliers. The movies were reviewed to visually assess the amount of motion and imaging artifacts in the data, and identify potential artifacts in the functional data which may not be apparent in the motion, GSC, or outlier timeseries.

Several automated QC summary measures were generated based on preprocessing outputs and related QC timeseries. These measures are described in Table 1 (QC summary measures section). Some of these measures provided an agnostic description of features of the original functional data, including the maximum value of GSC (**MaxGSchange**) and FD (**MaxMotion**). Since, often, these worst-case instances have already been identified as potential outlier scans, these measures inform about the state of the data prior to preprocessing. Other measures such as **MeanGSchange** or **MeanMotion** represent average GSC or FD values limited only to valid (non-outlier) scans, so they can be considered as more informative about the state of the data *after* preprocessing. Other useful statistics include the total number of run-specific outlier scans (**InvalidScans**), the number of non-outlier scans (**ValidScans**), and the proportion of valid scans (**PVS**), providing several indicators of the overall quality and amount of valid data within each individual run for each subject. Last, and aiming to directly quantify the performance of spatial *normalization* and its indirect effect on inter-modality coregistration, the measures **NORM_{func}** (functional normalization) and **NORM_{anat}** (anatomical normalization) measured the similarity between the gray matter mask in the normalized data and in a reference MNI atlas. Relatedly, **AFO** (anatomical to functional overlap) measured the similarity between gray matter masks in functional and anatomical images, evaluating the accuracy of inter-modality coregistration.

Participant-level denoising exclusion criteria included cases that were considered extreme in either the visual QC step, or in the automated QC summary measures. For automated QC summary measures, extreme values were considered those above the threshold $Q3 + 3 \text{ IQR}$ (or below $Q1 - 3 \text{ IQR}$, for those cases when extreme low values were indicative of problems in the data), where $Q1$ and $Q3$

represent, respectively, the first and third quartiles of the distribution of a measure across the entire dataset, and IQR represents their difference (inter-quartile range).

3.3.3. Quality control of denoised data

QC of denoised data aimed at evaluating the quality of the functional data after denoising. Since denoising is the last step when preparing the data before computing functional connectivity measures or performing other statistical analyses, quality control measures of the denoised data provide a way to globally evaluate the suitability of the resulting fMRI data for functional connectivity or other statistical analyses.

Participant-level visual QC aimed at evaluating possible patterns or other features that may be visible in the BOLD signal timeseries after denoising and which may be indicative of a possibly too liberal or too conservative denoising strategy. In particular, we reviewed run-specific plots rendering carpetplots (Power, 2017) of fully preprocessed BOLD timeseries before and after denoising, together with the traces of GSC, FD, and outliers timeseries. These were inspected to confirm that sudden and synchronized variations in signal intensity had been flagged as outliers, and that there are no visible residual large-scale patterns in the BOLD signal timeseries, which could indicate the persistence of global or widespread noise sources (for example, respiratory-related motion or artifacts can appear as patterns with frequency around 0.3 Hz). Carpetplots carry a rich set of information about the timeseries which, in combination with other indicators of potential problems in the data, allow researchers to hypothesize potential sources of noise that may be prevalent in the data, guiding the search of possible solutions.

Several QC summary measures were computed characterizing properties of the BOLD signal after denoising (Figure 1). These measures are described in Table 1 (QC summary measures section). The QC measure **DOF** computes the effective degrees of freedom of the BOLD timeseries after denoising. Lower values (close to zero or negative) indicate that denoising is overly aggressive for the number of functional scans available, and that noise correction comes at the expense of loss of meaningful variability severely impacting our ability to accurately estimate any model parameters of interest from the BOLD timeseries, such as functional connectivity measures or

task-related responses. The QC measure **BOLDstd** characterizes the stability of the BOLD signal after denoising. BOLDstd is a measure similar to MeanGSchange but computed from the data after denoising. It is inversely related to the BOLD signal temporal signal-to-noise ratio and, similarly to GCOR, high values are often indicative of the presence of potential noise sources in the residual fMRI data, although it needs to be interpreted with care as unusually low values can also indicate low effective degrees of freedom associated with the loss of meaningful variability from the BOLD timeseries. The QC measure **GCOR** (Global Correlation; Saad et al., 2013) represents the mean of functional connectivity measures (BOLD signal bivariate correlation coefficients) among all voxels, and it has been proposed as an effective control covariate for group-level analyses. GCOR often takes small positive values, caused by local correlations resulting in positive skewness in the distribution of functional connectivity values. High values can indicate an insufficient denoising strategy, and negative values can result from overly aggressive denoising, global signal regression, or biased-inducing denoising strategies.

Additional QC procedures and measures were derived from the distribution of functional connectivity (FC) values, computed as Pearson's r correlation coefficients between the BOLD signal timeseries after denoising among all pairs from a fixed set of 1,000 random voxels within the MNI-space gray matter template mask, in

order to evaluate a relatively dense sample of connections from the whole-brain connectome.

Visual inspection of these distributions allowed us to evaluate the relative presence of residual noise sources in the BOLD timeseries of each individual participant, which tend to shift the entire FC distribution toward positive values, altering the FC distribution center (representing the value GCOR) and its overall shape in a manner that is highly variable across different participants and across different runs. In comparison, the relative absence of noise sources is expressed as FC distributions that appear relatively centered (with a small positive distribution mean, and a distribution mode approximately at zero) and similar across different runs and participants.

Participant-level exclusion criteria included severe departures from expected FD distribution shapes after denoising – that is, with significantly skewed, shifted, flat, or bimodal distributions after denoising – as well as the presence of extreme outlier values in any of the computed QC measures (using the same $Q3 + 3 \text{ IQR}$ or $Q1 - 3 \text{ IQR}$ thresholds as before).

Last, the QC measure **QC-FC %** (percent match in QC-FC correlations) represents an individual quality control measure characterizing a property of the entire dataset, rather than properties of individual participants or runs. This measure is also computed from

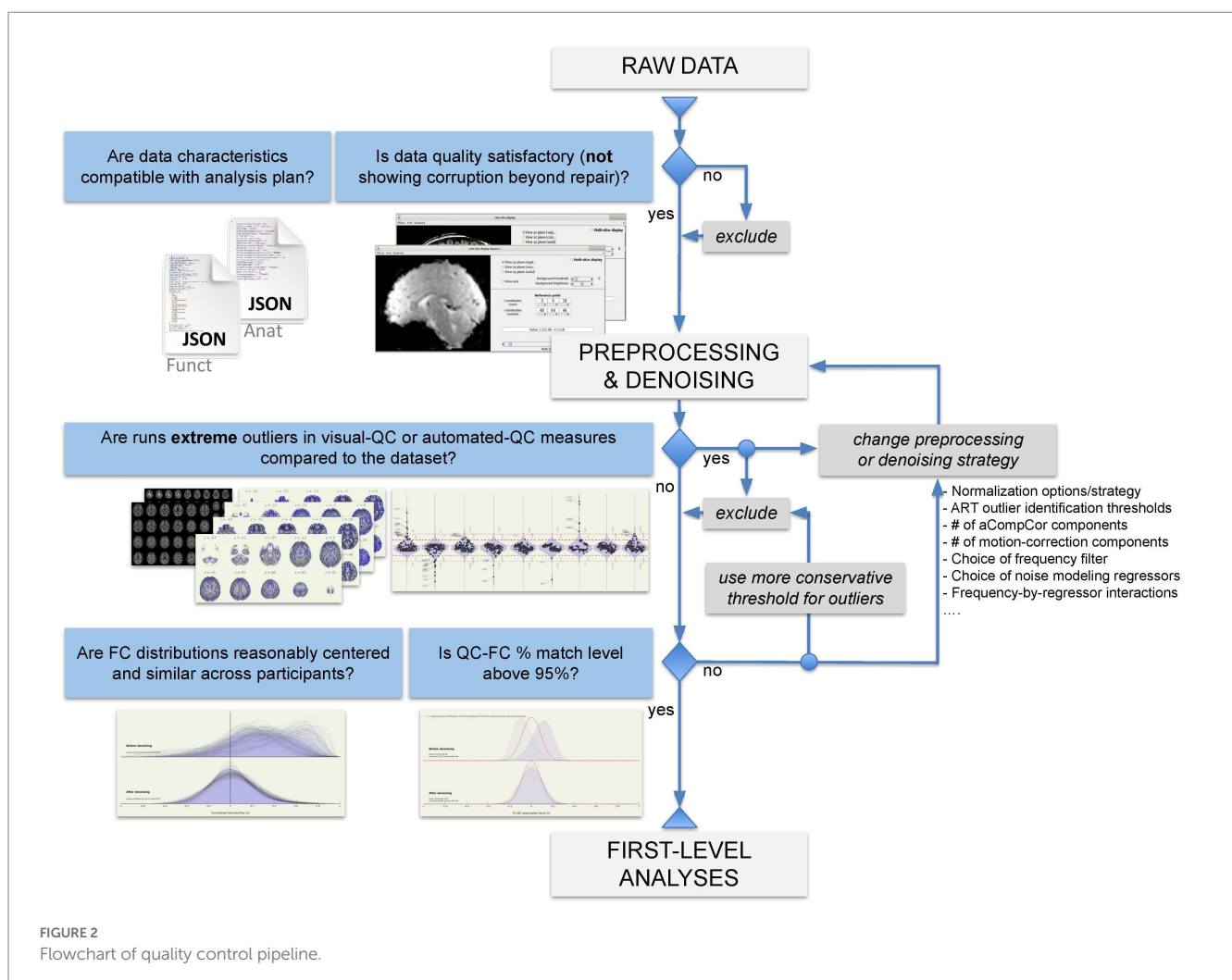


FIGURE 2
Flowchart of quality control pipeline.

these same distributions of FC values (one FC distribution per participant), but this time focusing on QC-FC inter-subject correlations (Ciric et al., 2017), evaluating whether changes in the spatial correlation structure of the BOLD data covaried with participant-level quality control measures. In particular, using the same sample of connections from the whole-brain connectome estimated in the FC distribution step above, we computed the bivariate Pearson's r correlation coefficients across participants between each of the estimated connectivity values and representative QC measures (MeanMotion, InvalidScans, and PVS). The resulting distributions of QC-FC correlations were evaluated to detect systematic biases by computing the distributional distance between these distributions and those expected by chance (in the absence of QC-FC correlations, as estimated using permutation analyses). QC-FC % values were used to evaluate whether the chosen combination of preprocessing and denoising steps, as well as the choice of thresholds for participant-level exclusion criteria and other QA procedures resulted in satisfactory fMRI data quality levels, and to choose between possible alternatives when necessary. Match levels above 95% were considered indicative of negligible modulations in the BOLD signal correlation structure, while lower values are considered indicative of the persistence of potential problems in the denoised data, requiring either alternative preprocessing and denoising choices or more severe participant exclusion criteria (Figure 2).

4. Results

4.1. Participants and data characteristics

Information reported here derive from investigating the nifti files characteristics directly or from their sidcar json files, which had been generated prior to release *via* unspecified procedures ($n = 124$) or *via* dcm2nii (Li et al., 2016) v1.0.20170314 ($n = 15$).

In this study, we analyzed resting state and anatomical MRI data from 139 participants acquired from 7 sites, including 151 functional runs and 139 anatomical images (mprage, 3D TFE, or unspecified). All sites contributed 20 participants except for site #3 ($n = 16$) and site #4 ($n = 23$). Throughout the manuscript, individual participants are referred to using both the collection's ID number (e.g., sub-____) where the first digit reflects the acquisition site of origin, and using ascending numbers (e.g., S____) representing participants ordered from site #1 to site #7.

The fmri-open-qc-rest collection was characterized by data with heterogeneous image resolution, scanner acquisition parameters, and experimental design. A detailed characterization of data features broken down by acquisition site is reported in Supplementary Tables S1, S2 for anatomical images, and in Table 2 for functional data.

Gathered information about functional data suggested that data were acquired by Siemens or Philips MRI scanners of various models (Trio Tim, Prisma Fit, Verio and Magnetom Trio, or Achieva or Achieva DS), using head coils with 12, 32, or unspecified number of channels. Data sampling differed on temporal (2- or 2.5-s TR) and spatial parameters, such as voxel dimensions (ranging from $1.6 \times 1.6 \times 3.1$ to 4 mm isotropic) and number of acquired slices (between 32 and 45). No information was available regarding whether any online processing was performed during or after acquisition, for example prospective motion correction or denoising. By design, the

experience of the participants was also different. Total time spent for the functional BOLD imaging acquisition ranged between 288 and 1,810 s (approximately between 5 and 30 min) which was acquired either in one continuous run or split into two ($n = 12$). During the functional data acquisition, participants were exposed to different visual stimuli (black screen with crosshair, eyes closed, or unspecified) and instructions (rest, relax and think of nothing particular, or unspecified).

Information incongruencies were encountered for sub-506 (S85) and sub-507 (S86) functional data, wherein 39 slice timings were reported in the sidcar json files but only 35 slices were available as per the nifti header information. This may suggest that these functional runs were not in a raw-level form or that the json files included faulty information.

There was no available information regarding several elements which had been shown to carry meaningful individual differences and which were relevant for data interpretation. No information was available regarding participant demographics (age, sex, medical and mental health history, mental and physical status at time of acquisition, psychoactive medication, etc.), participant inclusion and exclusion criteria, informed consent and assent. For example, the task description of sidcar json files of site #1 could be interpreted as suggesting that participants might include children who were asked to withhold taking psychostimulants the day prior to and the day of scanning; and the procedure description reported from the json files of site #5 could imply that participants were recruited under a study of brain traumas. Additionally, no information was available about the study paradigm, study design, or presence of experimental manipulation prior to or during data acquisition. Relatedly, it was not possible to determine whether the same individual was scanned in different sites or longitudinally, or if data were deemed unusable by the experimenters for any reason.

Critically, we did not know whether all or any of the above elements covaried with site and, consequently, whether potential inter-site variability encompassed meaningful individual differences in addition to heterogeneity associated with differences in scanner or acquisition details. Given the information available, or lack thereof, *site* was identified as a control variable. We cannot rule out that differences among sites may include meaningful factors, such as sample's age, health or medical status, or study design. These may legitimately affect BOLD signal properties of interest, including functional connectivity measures, in a manner that cannot be effectively separated from other sources of differences among sites, such as those resulting from differences between MR acquisition parameters or noise sources. Because of this, whenever possible we limited analyses of intersubject variability to focus only on within-site analyses, explicitly disregarding variability across sites due to the unavoidable issues when attempting to interpret sources of inter-site variability.

4.2. Raw-level data QC

Visual QC of the functional data identified different types of artifacts. We noticed artifacts appearing as spatial susceptibility distortion or signal drop out (e.g., sub-304 [S44]; Figure 3A), ghosting/aliasing (e.g., sub-717 [S136]; Figure 3B), signal inhomogeneity localized in regions of high tissue contrast [e.g.,

TABLE 2 Functional MRI data information for each acquisition site.

	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6	Site #7
N	20	20	16	23	20	20	20
Collection ID	Sub-101 to 120	Sub-201 to 220	Sub-301 to 316	Sub-401 to 423	Sub-501 to 520	Sub-601 to 620	Sub-701 to 720
CONN ID	S1 to S20	S21 to S40	S41 to S56	S57 to S79	S80 to S99	S100 to S119	S120 to S139
MRI scanner	Philips Achieva	Philips Achieva	Philips Achieva DS	/	Philips Achieva (5) Siemens Trio Tim (14) Siemens Prisma Fit (1)	Siemens Magnetom Trio	Siemens Verio
Head coil	/	/	32 channels	/	/	12 channels	/
Flip angle [°]	75	90	90	/	90 (17) 80 (3)	90	80
Phase encoding direction	j-	j-	j-	/	j- (15) / (5)	/	j-
Parallel acquisition technique	SENSE	SENSE	SENSE	/	/ (15) no_stimulation SENSE (5)	/	/
Voxel dimension [mm ³]	2.7×2.7×3 (19) 2.3×2.3×3 (1)	3×3×3.8	1.6×1.6×3.1	2.7×2.7×3	3×3×4 (15) 1.9×1.9×4 (5)	4×4×4	3×3×3.5
Field of view [slices]	96×96×47 (19) 112×112×47 (1)	80×80×38	128×128×45	96×96×47	80×80×35 (10) 128×128×34 (5) 80×80×34 (4) 80×80×39 (1)	64×64×32	64×64×39
Repetition time [s]	2.5	2	2.5	2.5	2	2.5	2.5
Acquired EPI runs	1	1	1	1	1	1 (8) 2 (12)	1
Scans acquired	156 (18) 128 (2)	150	162	123	144	[240–724]	198
Acquisition duration [s]	390 (18) 300 (2)	300	405	307.5	288	[600–1,810]	495
Slice timings available	Yes	Yes	Yes	/	Yes (13) / (5) wrong (2)	/	Yes
Task stimuli	White cross over black screen	Eyes closed	White cross over black screen	/	Eyes open	/	Eyes closed
Task instructions	/	Rest	Relax and think of nothing particular	/	/	/	/
Number of properties present in json file(s)	31	29	32	2	15 (5) 20 (13) 21 (2)	8 (8) 8 each run (12)	14

The information reported refers to all participants of each site, unless otherwise specified by the number in parenthesis reflecting the subset of participants. Participants are identified by the collection's ID number (e.g., sub-____) and by increasing numbers (e.g., S____) representing participants in ascending order. mm, millimeters; MRI, magnetic resonance imaging; properties of a json file, key-value pairs included in the json files; s, seconds; SENSE, sensitivity encoding; °, degrees; “/” indicates that information was not available.

sub-314 (S54); [Figure 3C](#)], of unspecified nature, or their combination [e.g., sub-409 (S65); [Figure 3D](#)]. For a complete list of identified artifacts broken down by participant and modality see [Supplementary Table S3](#).

Incorrect orientation of functional data was encountered for sub-518 (S97) and sub-519 (S98), which appeared upside-down. We considered to correct it by either applying a 180° rotation along the y-axis (i.e., preserving the relative position between the x, y, z axes) or a non-rigid reflection along the z-axis (i.e., flipping the data *via* a x, y, −z axis transformation which effectively would swap the signal between the left and right hemispheres). We opted to flip the data in both instances, based on the better visual match achieved between the flipped functional data and its respective anatomical data ([Supplementary Figure S1](#)).

During visual QC of anatomical data, we noticed few artifacts. Several participants from site #5 showed potential signs of past surgeries, as identified by localized darker areas (appearing as dots) traveling through contiguous slices reaching from the cortex to subcortical medial areas [e.g., sub-509 (S88); [Figure 4A](#), $z = 4$]. Often, these artifacts were localized in areas which appear to correspond to artifacts in the participant's functional data ([Supplementary Figure S2](#)). Sub-509 (S88) showed areas of intensity inhomogeneities bilaterally ([Figure 4A](#), $y = 5$ and $x = -35$) which appeared as bands in the y axis, and large asymmetrical lateral ventricles ([Figure 4A](#), $x = -17$). Other cases of potential anatomical variations or artifactual signal intensity were encountered including in sub-719 (S138; [Figure 4B](#)). Few cases of ringing-like patterns more prominently visible along the z-axis were noticed in a sub-218 (S38; [Figure 4C](#)) and in a few other anatomical

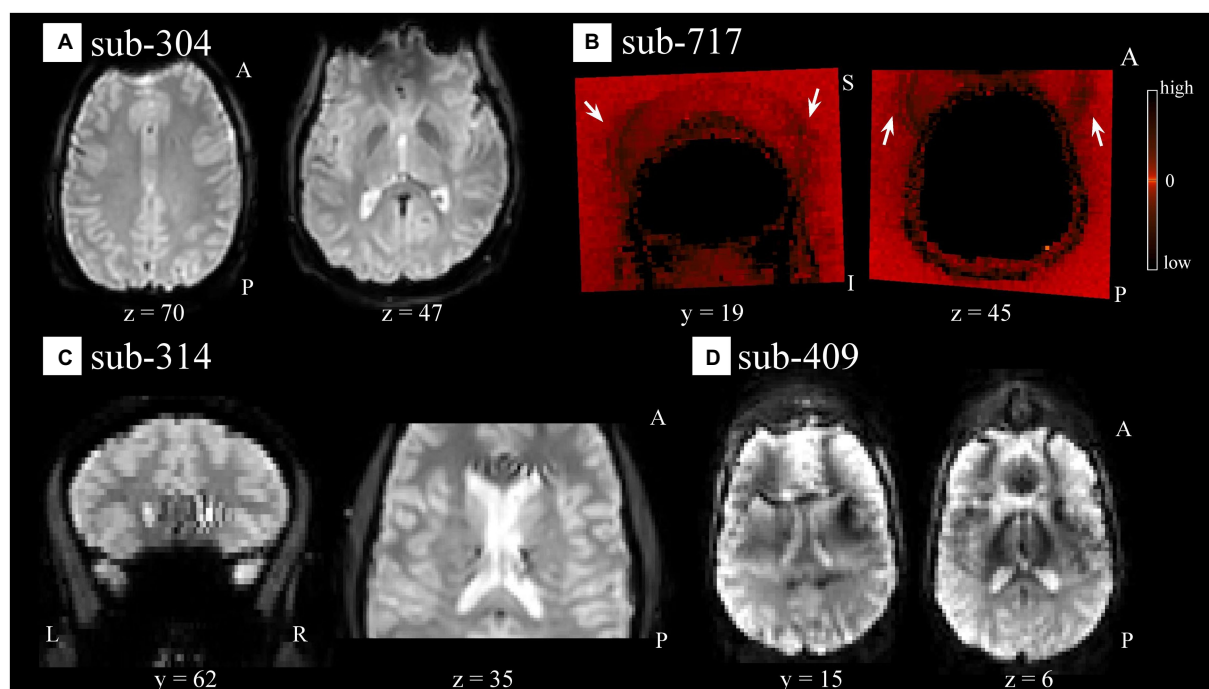


FIGURE 3

Spatial artifacts of raw-level functional data. **(A)** Spatial distortions and signal drop out in superior/orbito-frontal regions in sub-304 (S44). **(B)** Aliasing or ghosting in the coronal ($y=19$) and axial ($z=45$) slices from sub-717 (S136). For visualization purposes only, intensity values have been scaled so that low and high values would appear darker, making more evident artifacts such as those highlighted by the white arrows stemming from the superior (left image) and frontal (right image) regions of the head. **(C)** Unspecified signal inhomogeneity artifacts affecting sub-314 (S54) functional scans localized near areas of high intensity contrast such as CSF to WM. **(D)** Ghosting, spatial distortions, and signal inhomogeneities are noticeable in sub-409 (S65) functional data across all scans and several slices. For all panels, the images render the first functional scan of raw-level data.

images (see [Supplementary Table S3](#)). Additionally, there were few cases with noticeable motion-related and ghosting, of which sub-519 (S98; [Figure 4D](#)) was an example. Inasmuch the preprocessing of anatomical images for FC-MRI analysis was instrumental to preparing the functional data, a low(er) quality of anatomical images was not considered a major roadblock unless it produced a faulty segmentation or normalization.

During anatomical visual QC, we also observed what could be described as a skin marker on the forehead (right hemisphere) of most participants from site #5 ($n=15$) including all those scanned with Philips Achieva, and in a few from site #7. While there was no available information regarding which hemisphere the marker was placed on, and under the assumption that they would be placed in a standardized fashion, the consistent lateralization with which the marker was observed for all participants was considered as a hint of lack of left–right flip relative to one another.

Cross-modality visual comparison aided the characterization of artifacts. For example, unspecified signal intensity inhomogeneity was noticed in the functional data of sub-315 (S55; [Figure 5A](#), $x=2$), which corresponded to an undefined artifact or anatomical feature ([Figure 5B](#)). The artifact was localized in the medial-superior area above the cingulate cortex in the interhemispheric fissure, appearing dark in the functional data and bright in the anatomical images. Additionally, several examples of highly localized signal inhomogeneity with sharp intensity differences were characteristic of participants from site #5. From a visual inspection, those appeared similar to those reported in [Figure 5](#), but

the comparison with the anatomical data suggested that those could potentially derive from past brain surgeries (e.g., sub-509 [S88]; [Supplementary Figure S2](#)).

Overall, only one run corresponding to sub-409 (S65) was deemed to be excluded based on extreme spatial corruption severely affecting multiple slices and persistent across all scans. All other cases mentioned above were flagged as uncertain (see [Supplementary Table S3](#) for a complete list) as we considered that in the absence of additional indications their potential effect on the quality of the BOLD signal may not be severe enough to warrant exclusion.

4.3. Preprocessed data QC

Since fieldmaps were not available, our preprocessing included a direct, rather than indirect, normalization procedure to try to minimize EPI-specific warping caused by susceptibility distortions ([Calhoun et al., 2017](#)). Similarly, we skipped STC because slice timing information was available for only a portion of runs ($n=89$ out of 151) and most importantly, it was selectively missing for entire sites (#4, #6, and some cases from site #5). We elected to skip STC for all participants in order to prevent introducing variability driven by distinct analytical approaches into the results, which, in light of the characteristics of the fmri-open-qc-rest collection, could exacerbate potential inter-site (and in the case of site #5, even intra-site) heterogeneity even further.

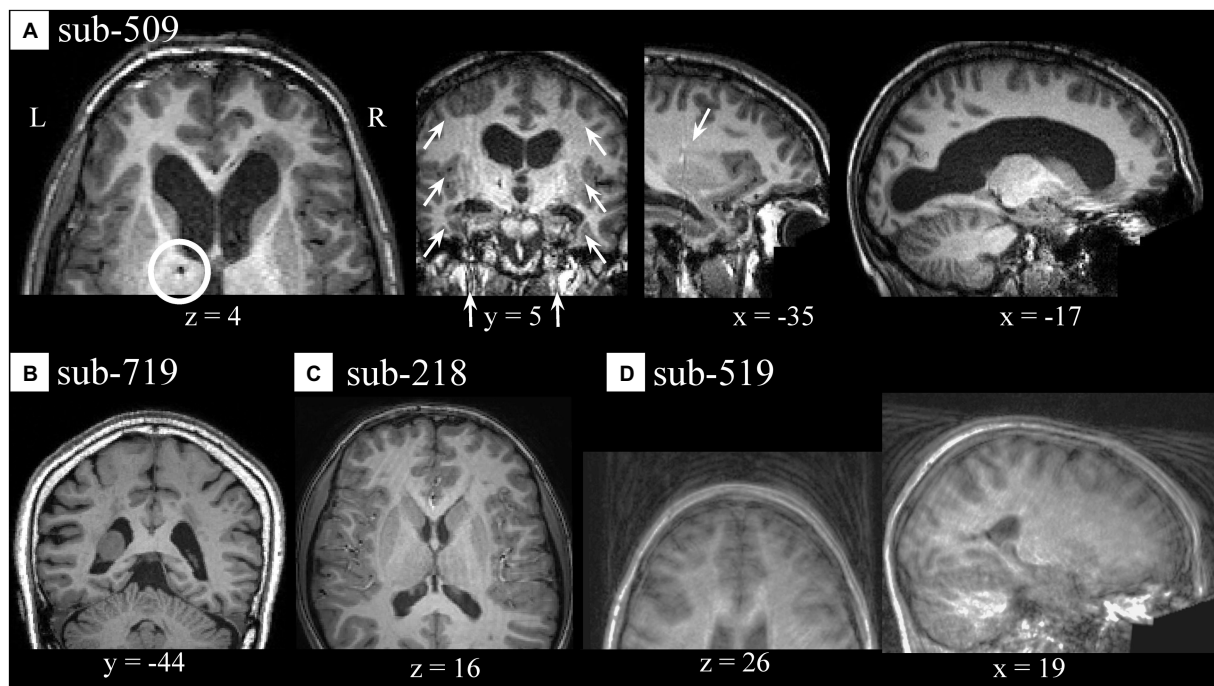


FIGURE 4

Spatial artifacts of anatomical raw-level data. (A) Sub-509 (S88) presented signs of potential past surgery ($z=4$) appearing as dark, small, localized areas traveling through several slices, signal intensity inhomogeneity localized bilaterally along the y -axis ($y=5$ and $x=-35$), and individual anatomical variations of size and shape of the lateral ventricles ($x=-17$). (B) Individual anatomical differences in the form of an asymmetrical mass or unspecified signal inhomogeneity localized in the lateral ventricle of a sub-719 (S138). (C) Motion-related artifacts or ringings in sub-218 (S38). (D) Sub-519 (S98) showed severe aliasing, ghosting, and/or motion-related artifacts.

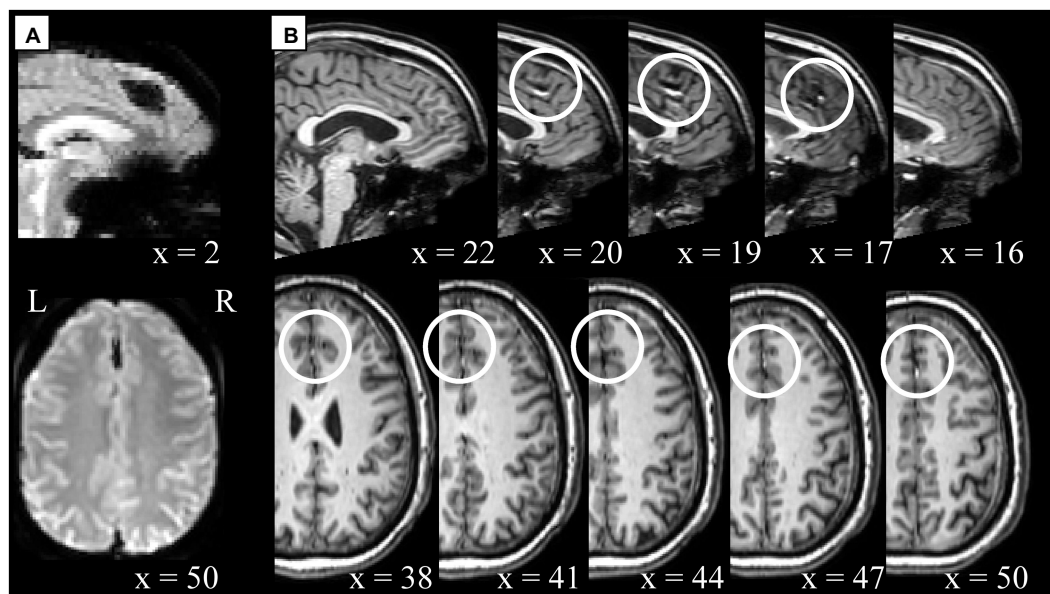


FIGURE 5

Example of cross-modality visual quality control for artifact characterization. Potential artifact of unspecified type in sub-315 (S55) functional (panel A) and anatomical (panel B) data. (A) Signal inhomogeneity affecting several axial slices localized in the interhemispheric fissure. The first scan is displayed here, however similar artifacts are noticeable across all scans. (B) Unspecified anatomical artifacts rendered in contiguous sagittal ($x=22$ to $x=16$) and axial slices ($x=38$ to $x=50$) in the top and bottom row, respectively. White circles indicate areas where artifacts are visible in a location comparable between functional and anatomical data. Note, the anatomical and functional images displayed here were in raw-level form, hence the spatial coordinates refer to subject-space and might differ across modalities.

Visual QC of preprocessed data identified severe failures of anatomical normalization and segmentation for sub-509 (S88) and sub-511 (S90). In both cases, the normalized anatomical and segmented tissue ROIs appeared fragmented and showed poor continuity within tissue type but sharp differences across tissues, see [Figure 6](#) (slices in row 7 columns 4 and 6) and [Figure 7B](#) (bottom).

Beyond those issues, visual inspection of the functional and anatomical data and potential residual artifacts in the functional timeseries identified no other obvious failures of functional preprocessing, including for the cases flagged as uncertain during raw-level data QC. For an overview of the full dataset after preprocessing, see [Figure 6](#) (anatomical images, $n = 139$) and [Figure 8](#) (functional scans, $n = 151$).

Automated QC measures (InvalidScans, PVS, MeanMotion, $NORM_{anat}$, $NORM_{func}$, and AFO in [Figure 9](#); other measures are reported in [Supplementary Figure S3](#)) were generated from $n = 151$ functional runs and $n = 139$ anatomical images ([Figure 9](#), left). Low extreme outliers (values 3 IQR below the 1st quartile) were identified for $NORM_{anat}$ [$n = 2$, sub-509 (S88) and sub-511 (S90)] and AFO [$n = 1$,

sub-509 (S88)], which corresponded to the cases identified during visual inspection. These data were also identified as extreme low outliers based on the distribution of total tissue volumes ([Supplementary Figure S4](#)). We visually inspected again the cases identified as mild low outliers from the distribution of $NORM_{anat}$ ($n = 2$; see sub-716 [S135] in [Figures 7A,B](#)), $NORM_{func}$ ($n = 0$), and AFO ($n = 0$) and confirmed that those indicated an acceptable preprocessing performance.

Several extreme low PVS outliers were identified ($n = 7$ with PVS below 75%): sub-118 (S18), sub-405 (S61), sub-519 (S98), sub-703 (S122), sub-706 (S125), sub-708 (S127) and sub-714 (S133) as well as several, mostly overlapping, extreme high InvalidScans participants ($n = 6$ with 48 or more InvalidScans): sub-519 (S98), sub-607 (S106), sub-703 (S122), sub-706 (S125), sub-708 (S127) and sub-714 (S133). The only participant with extreme high InvalidScans who did not have low PVS was sub-607 (S106), who, despite having 50 outlier scans, accounted for less than 7% of the total scanning session.

One participant [sub-111 (S11)] had a GCOR value (0.0534) borderline but below the level of extreme outlier (GCOR = 0.0535). However, this participant showed no obvious artifactual effects in

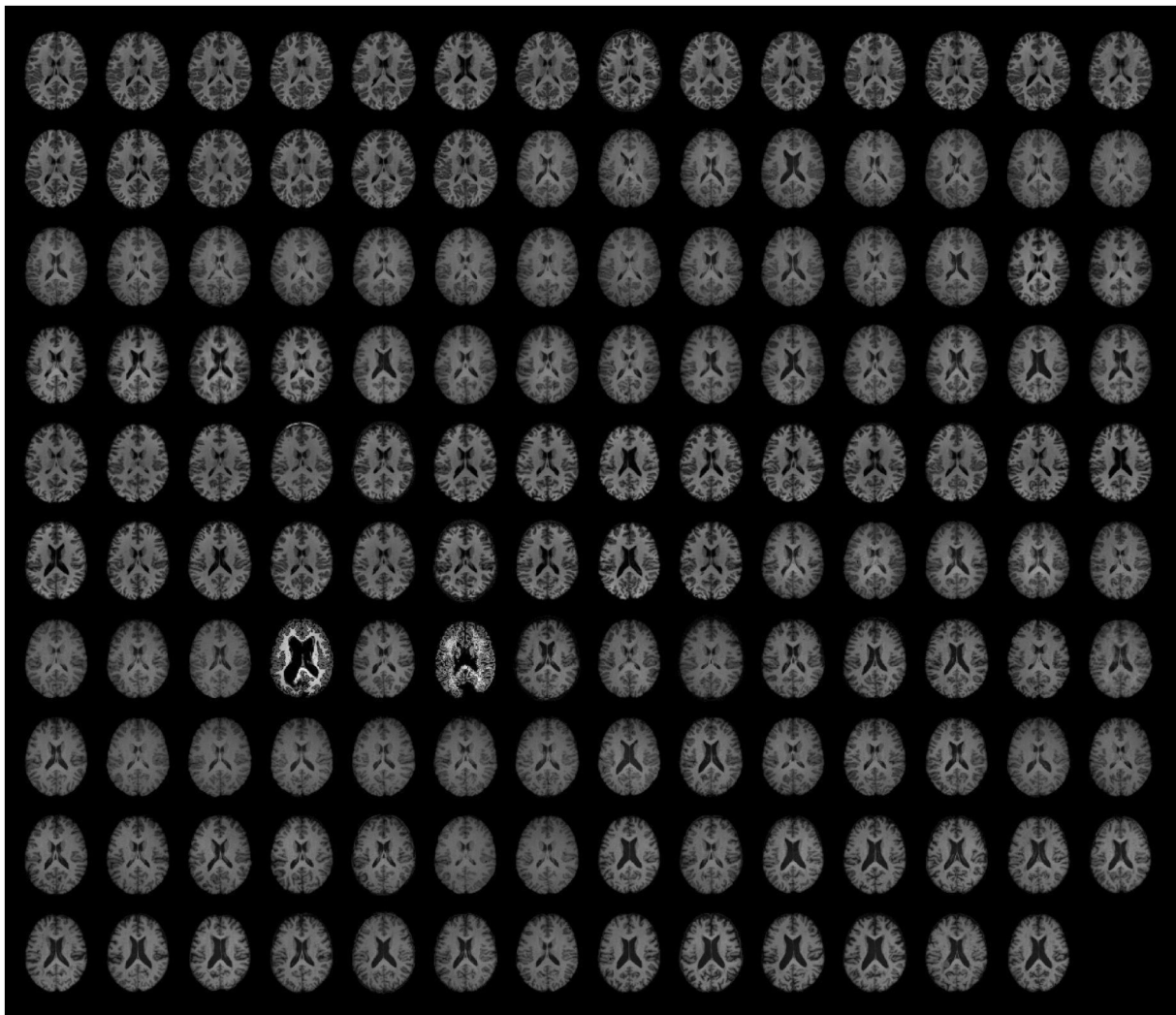


FIGURE 6

Preprocessed anatomical data. The same axial slice (MNI $z = 18$) of the fully preprocessed anatomical images is rendered for each participant ($n = 139$). For visualization purposes only, the BOLD signal intensity was scaled by the average value within each image.

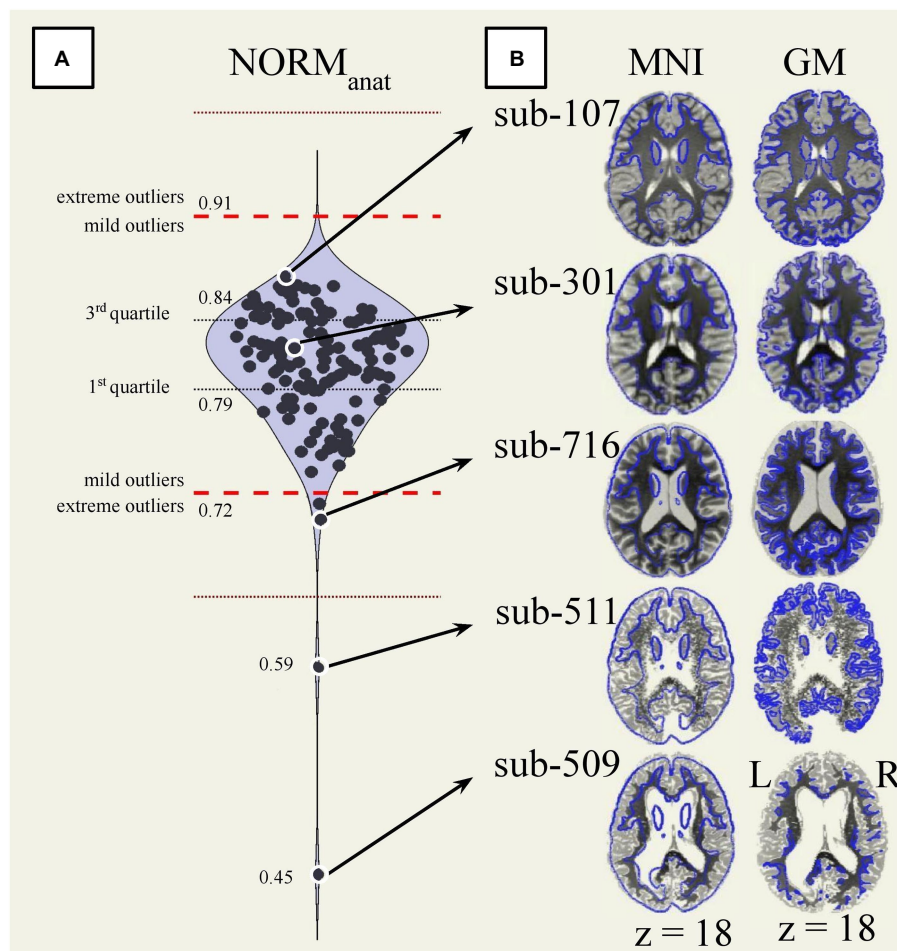


FIGURE 7

Automated and visual quality control of normalized anatomical data. **(A)** Distribution of the overlap between the normalized anatomical data ($n=139$) and the MNI-space ($NORM_{anat}$). Extreme outliers are identified as values 3 IQR above the 3rd quartile or below the 1st quartile (red dotted lines). Mild outliers are values 1.5 IQR above the 3rd quartile or below the 1st quartile (red dashed lines). **(B)** The same reference axial slice (MNI $z=18$) renders the normalized anatomical images from five participants. The participants' anatomical image is, on the left, overlaid on the 25% boundaries of the gray matter *a priori* probability maps MNI-space template (blue outline), and on the right, against each participant's anatomical gray matter boundaries. The participants reported in the figure are ordered from top to bottom based on their $NORM_{anat}$ values. Specifically, compared to the full dataset, sub-107 (S7) had the highest value, sub-301 (S41) was close to the median value, sub-716 (S135) was close to the low mild outlier threshold, sub-511 (S90) and sub-509 (S88) were the two lowest values and extreme outliers. GM, gray matter; MNI, Montreal Neurological Institute space; $NORM_{anat}$, overlap between the MNI-space and the normalized anatomical data.

carpetplots, or from other visual checks, nor had values in the mild (1.5 IQR) or extreme (3 IQR) outlier range for any other QC measure. Given that GCOR potentially includes some amount of meaningful intersubject variability, we elected not to exclude this run in order to avoid suppressing possibly natural variability.

Last, confirming our previous observations, there were strongly significant differences in all QC measures between the different sites (InvalidScans $F(6,132)=4.24$ $p=0.0006$, PVS $F=3.33$ $p=0.0044$, MeanMotion $F=8.85$ $p<0.0001$, $NORM_{anat}$ $F=13.22$ $p<0.0001$, $NORM_{func}$ $F=23.49$ $p<0.0001$, and AFO $F=13.42$ $p<0.0001$).

4.4. Denoised data QC

The distribution of automated QC measures (DOF, BOLDstd, and GCOR) for all denoised data ($n=151$ corresponding to 139 participants) is reported in Figure 9 (right). There were no extreme

outliers in BOLDstd, nor extreme low absolute DOF values, and participants with the lowest DOF values in this dataset [sub-519 (S98) DOF=17.1, sub-405 [S61] DOF=24.2, and sub-714 (S133) DOF=26.2] were already identified as extreme outliers with low PVS values. As with preprocessing QC measures, there were strongly significant differences in all QC denoising measures evaluated when compared between the different sites [DOF $F(6,132)=27.92$ $p<0.0001$, BOLDstd $F=19.65$ $p<0.0001$, and GCOR $F=12.98$ $p<0.0001$].

After preprocessing but before denoising, the distributions of functional connectivity estimates (FC distributions, Figure 10 left column) revealed severe biases, with connectivity values centered at $r=0.27$ on average across all participants, and also showed high levels of variability in the FC distribution center, with standard deviation 0.12 across participants. After denoising, the FC distributions (Figure 10, central column) were centered around $r=0.031$, and had low variability (standard deviation 0.01 across participants). Visually, FC distributions after denoising appeared

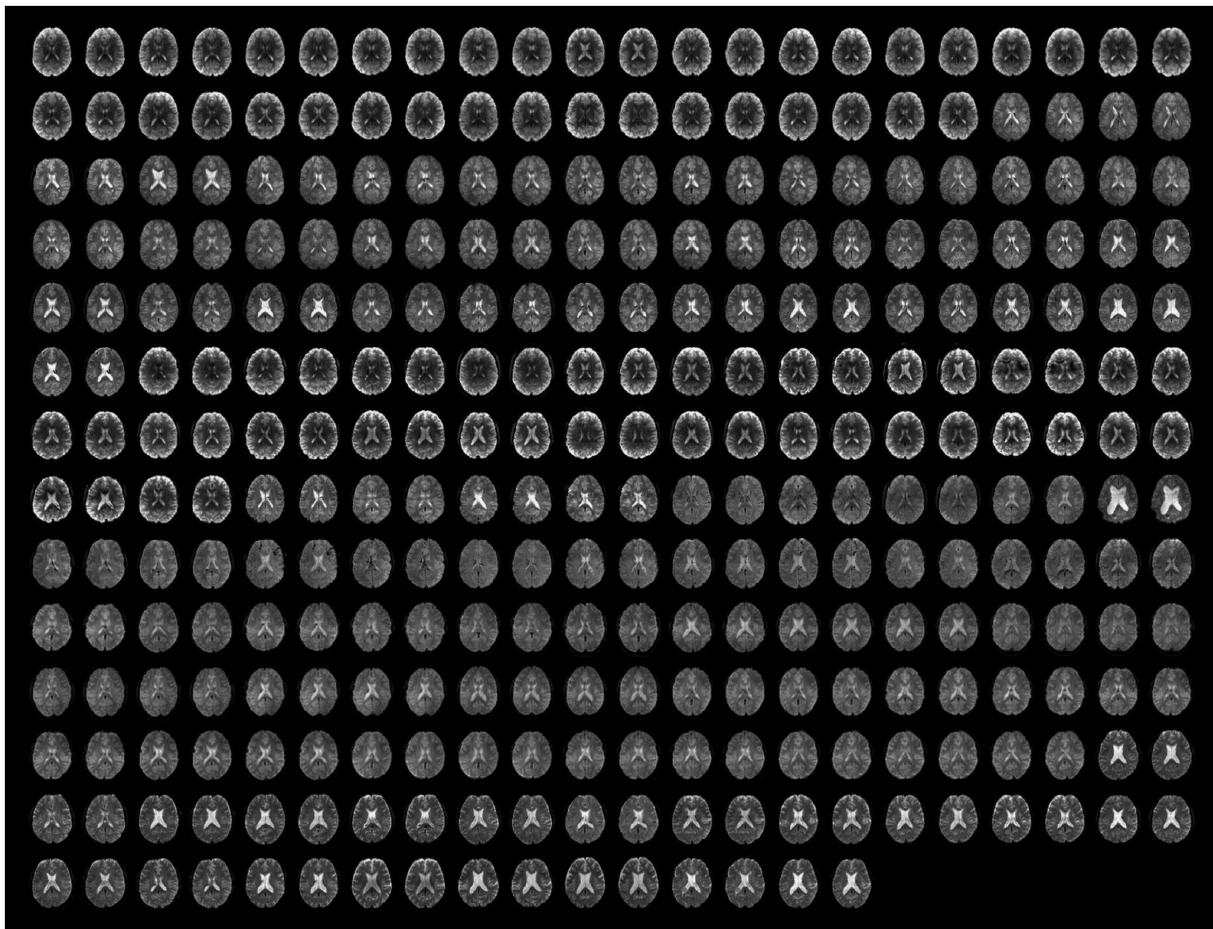


FIGURE 8

Preprocessed unsmoothed functional data. The same axial slice (MNI $z=18$) for the first and the last functional scan are rendered for all runs ($n=151$). For visualization purposes only, the BOLD signal intensity of each scan was scaled by its average value.

more centered and similar across participants, and nearly symmetrical with slightly longer positive than negative tails, as expected (for comparison, [Supplementary Figure S5](#) displays examples of FC distributions that could result if our denoising strategy had been overly or insufficiently aggressive in this same dataset).

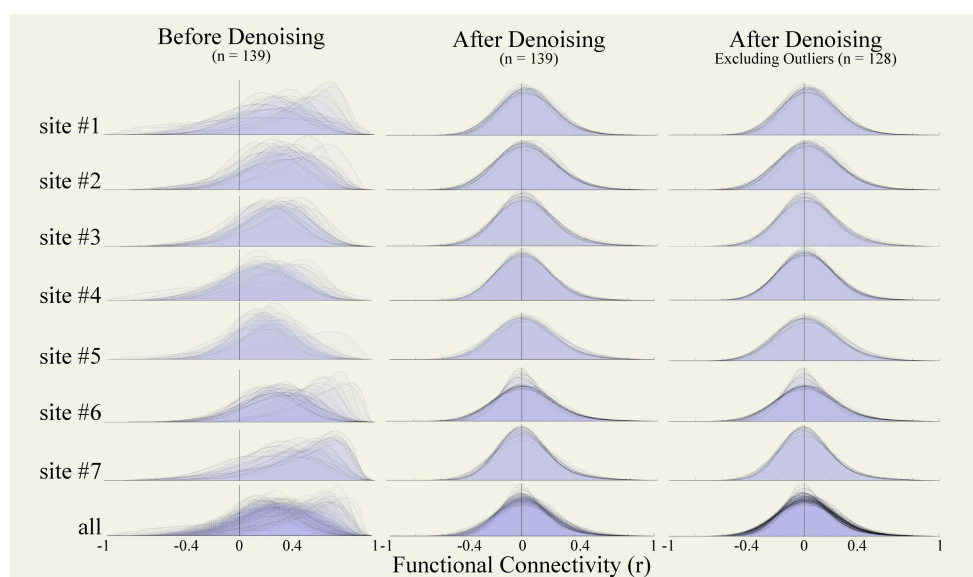
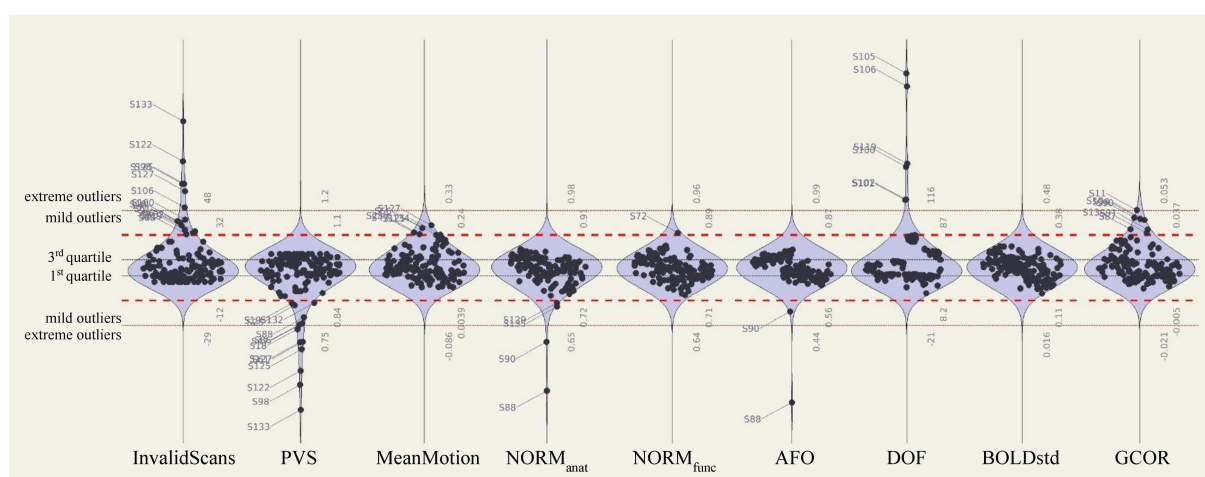
No individual runs were identified as potential outliers after denoising from visual inspection of these results. Site #6 included several runs with distinctive narrower FC distributions, but these were associated with scanning length that were considerably longer (identified in the [Figure 9](#) DOF distribution as having significantly higher degrees of freedom compared to other runs in this dataset). We did not exclude these runs but depending on the planned analyses it may be advisable to consider homogenizing the scanning duration length of the fMRI data.

QC-FC correlations were estimated separately within each site to avoid potential site confounder effects. Before denoising, QC-FC correlation distributions showed poor percentage match levels, indicating the persistence of motion and data quality effects on functional connectivity estimates after preprocessing. Specifically, percentage match levels were below the 95% cutoff for InvalidScans [average within-site %match = 86.70 ± 11.77 ranged (65.82; 97.59)],

MeanMotion [85.37 ± 13.94 (56.78; 98.52)], and PVS [83.70 ± 11.51 (65.82, 97.59)], see [Figure 11](#) (left) and [Table 3](#) (top).

Denoising increased the percentage match levels of QC-FC distributions ([Figure 11](#) middle and [Table 3](#) middle) for InvalidScans [average within-site % match = 94.24 ± 2.56 (91.47; 97.68)], MeanMotion [96.82 ± 1.07 (95.64; 98.89)], and PVS [94.21 ± 2.50 (91.47; 97.26)]. Despite this, several QC-FC correlations still did not pass the desired 95% cutoff for at least one of the three evaluated QC measures, including site #3, site #4, site #5, and site #7 ([Table 3](#)).

Excluding all runs with identified extreme outliers in any of the evaluated QC measures ($n = 10$, 1 run identified during raw-level visual QC, 2 runs with problems in spatial normalization, and 7 runs with extreme low PVS) increased the percentage match level of QC-FC distributions for InvalidScans [average within-site % match = 96.79 ± 2.07 (92.35; 98.48)], MeanMotion [97.64 ± 1.03 (96.12; 99.21)], and PVS [96.75 ± 2.04 (92.35; 98.48)]. Despite this, QC-FC correlations of site #3 still did not pass the desired 95% cutoff. Since the distribution of PVS did not show a clear cutoff among those participants with extreme outliers and those with mild outliers, we decided to re-evaluate QC-FC correlations varying the PVS threshold used for participant-level exclusion, excluding one



desired 95% threshold in QC-FC match levels across InvalidScans [97.3 ± 0.89 (95.92; 98.48)], MeanMotion [97.77 ± 0.86 (97; 99.21)], and PVS [97.26 ± 0.86 (95.92; 98.48)], see [Figure 11](#) (right column) and [Table 3](#) (bottom). Automated QC measures of the final $n=11$ excluded participants and their carpetplots are reported in [Supplementary Figures S6](#) and [S7](#), respectively.

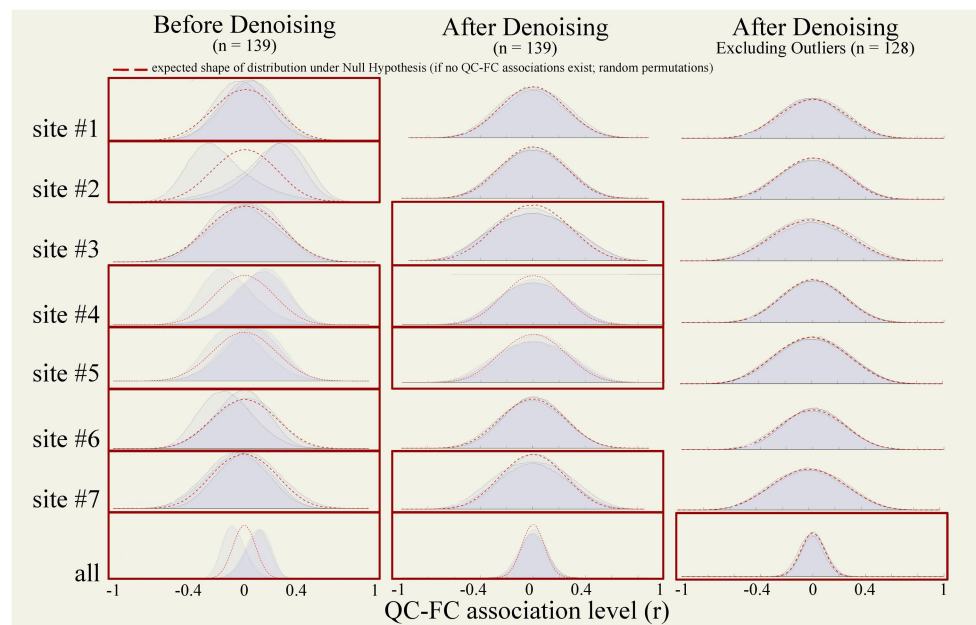


FIGURE 11

QC-FC correlation distributions. QC-FC plots tested functional connectivity associations with three nuisance factors (MeanMotion, InvalidScans, and PVS). Plots were generated from functional data from all participants ($n=139$) before (left) and after (middle) denoising, and after excluding outlier runs (right) identified during raw-level, preprocessed, and denoised data QC ($n=128$). Analyses were performed within each site independently (top) and across all sites jointly (bottom row). Red boxes indicate QC-FC with at least one QC-FC distribution that did not reach above the 95% cutoff. Red dotted lines represent a theoretical artifact-free null-hypothesis distribution. QC, quality control; FC, functional connectivity.

5. Discussion

In this study, we presented the CONN quality control pipeline (Table 4; Figure 2) based on a combination of visual and automated QC procedures. Publicly available resting state data were analyzed to showcase a complete QC workflow for the screening of raw-level, preprocessed, and denoised data for volume-based FC-MRI analysis. This pipeline includes visual-QC steps, where researchers visually judge the severity of potential artifacts in the raw, preprocessed, and denoised data, as well as a number of automated QC measures quantifying relevant aspects of the functional data. We recommend that researchers use the combination of visual- and automated- QC measures to motivate possible changes in their data preprocessing or denoising strategy that would address the issues raised by these measures, or, ultimately, to determine a list of individual participants or runs that may need to be excluded from the main analyses. The choice of a threshold for participant exclusion should be informed by the characteristics of one's own sample and the needs of their research questions or planned analyses. Rather than using absolute thresholds in QC measures, we suggest that sample-specific thresholds, such as the choice of a classical “extreme outliers” threshold of $Q3 + 3 \text{ IQR}$ for extreme high values (or $Q1 - 3 \text{ IQR}$ for extreme low values) are a reasonable starting point for participant exclusion. Last, our QC workflow uses the measure QC-FC %, characterizing the presence of inter-subject associations between functional connectivity and subject motion or outlier prevalence, and the stability of the FC distributions across different runs and participants ($FC \text{ mean} \pm SD$), as a way to evaluate the overall quality of the data, helping guide possible choices between alternative preprocessing and denoising strategies or participant exclusion thresholds.

Our QC workflow included a combination of procedures, of which some can be quantified precisely and even automated, while others cannot and will ultimately rely on each researcher's experience and judgment. In both cases, our approach is not that there is an “optimal” or even “correct” form of QC, but rather to encourage researchers to understand the rationale behind performing QC, follow a reasonable set of procedures, justify their choices during QC, and report their decision process when sharing their results to the community. For example, there is currently no agreed-upon correct choice or criterium of what constitute severe ghosting or other image artifacts, but our recommendation is for researchers to perform visual QC to evaluate the presence and severity of artifacts in their data, and then to define, based on their own criteria, experience, research goals, and specificities of their sample, what constitutes possibly extreme cases that would justify their exclusion. From this general perspective, we have attempted to provide specific measures and thresholds that could be used as precise exclusion criteria when possible (as sample-specific outliers, using a $Q3 + 3 \text{ IQR}$ threshold for individual QC measures, and as an absolute 95% threshold in QC-FC percent match levels), while also leaving room for other less easily quantifiable aspects of QC (using severity scores based on a researcher's own criteria during visual QC, and judging the overall level of centering and similarity of the QC distributions across the different subjects in our sample).

In that context, several automated QC measures were proposed to aid the identification of potential problems in the data or faulty preprocessing. $NORM_{\text{anat}}$, $NORM_{\text{func}}$, and AFO measures can be useful to evaluate functional normalization, anatomical normalization, and between modality coregistration success. Similarly, the relative severity of participant motion and other events

TABLE 3 FC density distributions and QC-FC correlations.

Site	<i>n</i>	<i>n</i> excluded	FC mean \pm SD	InvalidScans-FC	MeanMotion-FC	PVS-FC	QC-FC performance
Before denoising (<i>n</i> = 139)							
Site #1	20	/	0.27 \pm 0.13	90.81	91.38	92.01	Below cutoff
Site #2	20	/	0.29 \pm 0.08	65.82	56.78	65.82	Below cutoff
Site #3	16	/	0.24 \pm 0.07	97.59	98.52	97.59	
Site #4	23	/	0.17 \pm 0.08	75.85	78.41	75.85	Below cutoff
Site #5	20	/	0.2 \pm 0.07	87.58	91.03	87.58	Below cutoff
Site #6	20	/	0.32 \pm 0.13	97.46	91.71	75.24	Below cutoff
Site #7	20	/	0.39 \pm 0.11	91.78	89.76	91.78	Below cutoff
All	139	/	0.27 \pm 0.12	55.18	58.83	64.64	Below cutoff
After denoising (<i>n</i> = 139)							
Site #1	20	/	0.04 \pm 0.02	95.73	98.89	95.96	
Site #2	20	/	0.04 \pm 0.01	97.19	97.02	97.19	
Site #3	16	/	0.03 \pm 0.01	92.35	96.12	92.35	Below cutoff
Site #4	23	/	0.02 \pm 0.01	93.01	96.07	93.01	Below cutoff
Site #5	20	/	0.03 \pm 0.02	92.27	95.64	92.27	Below cutoff
Site #6	20	/	0.03 \pm 0.01	97.68	97.04	97.26	
Site #7	20	/	0.02 \pm 0.01	91.47	96.99	91.47	Below cutoff
all	139	/	0.03 \pm 0.01	91.41	94.20	90.02	Below cutoff
After denoising and excluding outliers (<i>n</i> = 128)							
Site #1	19	1	0.04 \pm 0.02	96.40	97.60	96.48	
Site #2	20	0	0.04 \pm 0.01	97.19	97.02	97.19	
Site #3	15	1	0.03 \pm 0.01	95.92	97.00	95.92	
Site #4	21	2	0.02 \pm 0.01	98.00	98.01	98.00	
Site #5	17	3	0.03 \pm 0.02	98.48	98.50	98.48	
Site #6	20	0	0.03 \pm 0.01	97.68	97.04	97.26	
Site #7	16	4	0.02 \pm 0.01	97.47	99.21	97.47	
All	128	11	0.03 \pm 0.01	97.24	95.84	93.97	Below cutoff

Values reported under FC mean represent the average \pm standard deviation across participants of GSC, the mean values of the FC density distributions, and QC-FC represent the percentage match level values, characterizing the presence of inter-subject associations between functional connectivity and subject motion or outlier prevalence. Bold font indicates % match values that are above the 95% cutoff. QC-FC performance values indicate whether any QC-FC measure percentage match level is below the 95% cutoff. FC, functional connectivity; GSC, global signal change; PVS, proportion of valid scans; QC, quality control.

that may cause outliers in the scan timeseries can be quantified using measures such as average of framewise displacement (MeanMotion), and the number or proportion of identified outlier scans (PVS). Measures evaluating the effective degrees of freedom of the BOLD signal timeseries after denoising (DOF), as well as its variability and intercorrelation (for example BOLDstd and GCOR), can also be useful to identify potential problems in the BOLD signal of individual participants before proceeding to statistical analyses. As other QC measures computed after preprocessing and denoising, outlier values in these measures may depend on the combination of most analytical steps that preceded it, so they do not directly suggest a potential source or cause of the identified problems. Finally, QC-FC correlations evaluate whether changes in the spatial correlation structure of the BOLD data covaried with participant-level quality control measures, such as the extent of participant motion, and the number or proportion of outlier scans, so they can be used as general measures of data quality to guide other data processing choices.

In this dataset these measures were used to evaluate the quality of the fMRI data and help guide our choices of denoising and exclusion procedures. Altogether, the QC pipeline and exclusion criteria adopted (Table 4) excluded 8% of the participants and minimized the presence of a variety of noise sources in the data as evaluated using a combination of visual and automated QC measures and procedures.

Many reasons may explain why bias persists after a successful preprocessing and adequate denoising, and these reasons create a multi(uni)verse of effective possibilities to counteract. Although relevant to the understanding of QC procedures, the evaluation of different processing pipelines was outside the scope of this paper and has been discussed in several seminal papers about preprocessing (Friston et al., 1996; Strother et al., 2004; Murphy et al., 2009; Chai et al., 2012; Hallquist et al., 2013; Power et al., 2014; Ciric et al., 2017) and denoising strategies (Churchill and Strother, 2013; Parkes et al., 2018; Maknojia et al., 2019; Tong et al., 2019; De Blasi et al., 2020; Golestani and Chen, 2022; for a review, see Caballero-Gaudes and Reynolds, 2017).

TABLE 4 CONN quality control pipeline checklist and exclusion criteria for whole brain resting state functional connectivity analysis.

	Category	QC Checklist	Tools	Exclusion criteria
Raw-level data	Source of heterogeneity of no interest (defined by the data intended used)	Acquisition parameters	MRI data	(A) Data that do not meet criteria for the specific analysis goals as defined by each individual research study
		Demographic	Sidcar json files	
		Task design	Scan sequences protocol	
	Artifacts	Ghosting	Visual inspection (scan-to-scan and slice-to-slice)	(B) Data corrupted beyond repair as judged by rater
		Aliasing		
		Foreign objects artifacts		
		Dropouts/truncation		
		Ringing		
		Spatial distortions		
		Contrast inhomogeneities		
	Personalized preprocessing needed	Artifacts that may require personalized consideration	Visual inspection (slice-to-slice)	
	Challenging data features	Motion related artifacts Anatomical variations	Visual inspection (scan-to-scan and slice-to-slice)	
Preprocessing	Failures of functional preprocessing	Artifacts in the timeseries	Visual comparison between the scan-to-scan movie of a reference functional slice with motion, GSC, and outlier timeseries traces	
		Normalization	Visual comparison between normalized functional data and MNI template	(C) † Functional data which cannot be preprocessed satisfactorily as judged by rater
			Visual comparison between anatomical gray matter and normalized functional data	
			Automated QC measure $NORM_{func}$	(D) † Cases with extreme values, as judged by a sample-specific Q1-3 IQR threshold criterion
	Failures of anatomical preprocessing	Normalization and segmentation	Visual comparison between normalized anatomical data and MNI template	(E) † Anatomical data which cannot be preprocessed satisfactorily as judged by rater
			Visual comparison between anatomical gray matter and normalized anatomical data	
			Automated QC measures AFO and $NORM_{anat}$	(F) † Cases with extreme values, as judged by a sample-specific Q1-3 IQR threshold criterion
Denoising	Residual noise factors	Within-participant	Visual comparison of carpetplots with motion, GSC, and outlier timeseries traces	
		Between-participant	Other QC variables: distribution of participant-level QC measures	(G) † Cases with extreme values in PVS, MeanMotion, or DOF, as judged by a sample-specific Q3 + 3 IQR or Q1-3 IQR threshold criterion
			Distribution of functional connectivity values	(H) † Extremely skewed, shifted, flat, or bimodal functional connectivity distributions after denoising, as judged by rater.
				Also used to guide preprocessing, denoising, and participant-exclusion-criteria choices.
			Distribution of QC-FC associations, for InvalidScans, MeanMotion, and PVS	Used to guide preprocessing, denoising, and participant-exclusion-criteria choices.

Cases with extreme values could be represented by values below 3 times the interquartile range above the 3rd quartile or below the 1st quartile, depending on the specific QC measure, compared to the full dataset distribution. BOLD, blood oxygenation level-dependent; FC, functional connectivity; GCOR, global correlation; MNI, Montreal Neurological Institute; QC, quality control; TPM, tissue probability map. † Indicates exclusion criteria applied only if potential remedial analytical or processing alternatives fail.

It is nevertheless important to note that not all measures that are used to evaluate the quality of the fMRI data in the context of QC procedures can or should be used to compare different preprocessing or denoising pipelines. In general, global or sample-level properties such as QC-FC %, characterizing between-subject QC-FC correlations, and FC mean \pm SD, characterizing between-subjects variability in the shape of FC distributions, are meaningful measures that can be used to guide choices in preprocessing and denoising, and in particular to compare the relative success of different preprocessing pipelines. In contrast, many measures, such as BOLDstd, DOE, MeanGSChange, which are designed to provide useful contrasts when comparing different participants undergoing the same acquisition and analytical procedures, should be considered with extreme care in the context of comparing different analytical procedures or pipelines, as they provide only a very limited view of the overall quality of the data, with often contradictory results when interpreted as direct measures of data quality.

We encourage researchers to consider preprocessing and denoising strategies as an array of tools to use on their data, and rely on quality control measures described above to help guide and substantiate their choice of the best tools to use for each dataset. Indeed in our case, QC testing did suggest to evaluate alternative analytical approaches to attempt to improve the overall quality of the results. For example, there were two cases [sub-509 (S88) and sub-511 (S90)] in which anatomical normalization failed. This could have suggested that trying alternative normalization procedures customized to the dataset could have been tested. For example, normalization approaches using lesion-informed templates (which could have been relevant for site #5), age-specific normalization templates, or different normalization parameters could have led to overall better normalization performance for these two cases and perhaps others. Moreover, we did not perform STC to avoid introducing artificial heterogeneity between and within sites driven by differences in preprocessing pipelines. Our choice was based on a lack of information regarding slice timings for a portion (41.6%) of the data. But in a real-life context, we would have reached out to the research groups where the data originated trying to find said information. Similarly, we would have reached out to the site#5 to confirm that sub-518 (S97) and sub-519 (S98) functional data needed to be flipped rather than rotated. Also, the QC-FC 95% benchmark was not reached for PVS when considering data from all sites jointly (Figure 11, bottom row). That indicates that if we want to perform analyses jointly across all sites, we would need to correct site effects, as those potentially contain a mixture of noise sources together with perhaps other meaningful differences in sample demographics, but similarly other site homogenization approaches could be attempted to try to reduce or remove the residual QC-FC correlations across sites. In deciding the best course of action for the fmri-open-qc-rest collection, we faced a tradeoff between maximizing power (i.e., including as much data as possible) and prioritizing the optimal approach for the majority – but perhaps not the totality – of the data. Excluding a portion of runs ($n = 11$ out of 151 runs, corresponding to $n = 11$ out of 139 participants) resulted in an overall more lenient approach to the rest of the data and minimized the estimated residual bias driven by invalid scans, proportion of valid scans, and mean motion within each site independently and improved it across all sites jointly. Ultimately, the data and the research question motivating one's own analysis will define what the “best” approach entails, potentially involving different analytical strategies. Whichever that is, we stress how

reporting the rationale guiding preprocessing and denoising choices in a study and supporting those choices with reports describing the associated QC measures and procedures used, is a key element for results interpretation and reproducible science.

The proposed QC workflow, checklist, recommendations, and exclusion criteria are agnostic of the analytical software employed. While designed and discussed around the implementation in CONN, our recommendations generalize to data fully or partially analyzed (preprocessed and/or denoised) *via* other software packages including AFNI (Cox, 1996), SPM (Friston and *Al*, 2007), FSL (Jenkinson et al., 2012), FreeSurfer (Fischl, 2012), fMRIPrep (Esteban et al., 2019), Tedana (DuPre et al., 2021), MRIQC (Esteban et al., 2017), pyfMRIQC (Williams and Lindner, 2020), and others. For example, $NORM_{anat}$, $NORM_{func}$, and AFO are measures diagnostic of preprocessed data quality, but they can be computed independently of the software or process that generated them. Furthermore, while the analytical details used to generate well-known metrics (framewise displacement, CompCor components, etc.) or methods (ICA, AROMA, CompCor) may vary across software packages, we expect that the recommendations provided in this manuscript should generalize beyond the specific measures used in the example presented in this manuscript. For example, we have no reasons to believe that the data exclusion based on the extreme departures of PVS relative to the sample's distribution should be specific to the outlier threshold or motion estimation method that we used, rather they could generalize to alternative definitions of FD (Jenkinson et al., 2002; Power et al., 2012). In a similar fashion, considerations about visual QC could be expanded to apply to data inspected through MRI image viewers or visual plots generated with alternative methods.

The FMRI Open QC Project dataset (Taylor et al., 2022) combines information from multiple sites. The preprocessing, denoising, and QC steps discussed in this manuscript did not directly address the issue of data harmonization across sites (Friedman et al., 2006; Yu et al., 2018). Effective harmonization of features across sites would require a considerably richer array of information from the sampled participants in order to be able to differentiate among intersite differences that may carry meaningful information, such as those due to differences in age and health status of participants sampled in different sites or studies, from intersite differences that may be related to other factors of no interest, such as those introduced by specific acquisition details used in each study. Despite this, the quality control procedures described in this manuscript attempted to focus, whenever possible, on features of the entire dataset, treating *site* as one would normally treat different subject groups in a single-site study, except for QC-FC correlations, where we chose to focus only on intrasite analyses as otherwise the results would be naturally confounded by some of the very large differences in QC measures observed among sites. QC procedures in the context of multisite studies would benefit from an integrated approach to data homogenization and quality control, which is still an open area of research.

Most of the QC pipeline that we had described for resting state functional connectivity analysis is also suitable for task-based connectivity and task-based activation analyses. The QC workflow and exclusion criteria related to raw-level data visual inspection, preprocessed data visual and automated procedures (e.g., $NORM_{anat}$, $NORM_{func}$, AFO, and PVS) apply to (f)MRI data regardless of the final intended analysis goal. However the nature of the analysis (connectivity vs. activation) and of the behavioral/cognitive processes elicited during data acquisition (to rest or to perform an explicit task) carry distinct potential dangers on the final statistical analyses and require customized considerations. For example, motion is highly problematic for functional

connectivity analysis, as it introduces biases reducing the accuracy of results, so it is thus usually more aggressively controlled for in the context of resting state analyses. In contrast, in task-activation studies, this is usually less of a concern as motion tends to simply reduce power (i.e., lowering statistical significance of the results) rather than introducing spurious results. Yet, activation analysis could suffer from a similar curse when motion artifacts are unbalanced between task conditions (e.g., larger subject motion during rest blocks compared to task blocks), so in the context of task-activation analyses QC measures that focus on the presence of task-correlated motion are often recommended. While the general QC workflow described in this manuscript can be equally used in the context of task-activation or other types of analyses, we would expect that the inclusion of additional QC measures focusing on analysis-specific features or sources of concern (e.g., quantifying the presence of task-correlated motion or other task-correlated noise sources in the context of task-activation analyses) would be necessary in order to better capture the suitability of the resulting data for those specific analyses.

Overall, the guidelines of our QC approach were to improve data quality and quantify residual nuisance effects. However, these guidelines were constrained by at least four limitations, which are the objective of open and active lines of work in the neuroimaging field. First, the field currently lacks a ground truth of what the BOLD signal is. It follows that quantifying the differences between the actual signal and the true signal was limited in its scope. Second, neural and non-neural signals are best thought of as a continuum rather than two ontological classes. Although regarded as a viable approach to minimize well-known bias, regressing out “non-neural” components might also have removed neural signals too (for example see Wang et al., 2021). Third, we applied similar processing to all data regardless of specific acquisition parameters, but it has been shown that non-harmonized MRI data could introduce spurious heterogeneity in FC estimates. However, potential sources of heterogeneity (e.g., inter-run, inter-participant, and inter-site variability; Greve et al., 2012) may be intertwined with true individual differences. Considering all available data, hence maximizing power and heterogeneity, may promote generalizability and reproducibility of neuroimaging results. Lastly, we defined exclusion criteria and cutoffs based on relative terms rather than absolute, which risks leading further away from a standardization of QC procedures. However, we argue that this shortcoming not only provides a necessary level of flexibility in view of the heterogeneity in acquisition details, sample characteristics, and experimental designs across different studies and fields, but also that it might effectively be overcome if QC procedures were to be consistently reported alongside FC results, however varied the QC strategies may be. Similarly to how distinct analytical approaches are regarded as equally valid in addressing the same research questions (Botvinik-Nezer et al., 2020), different QC pipelines could represent effective alternatives. As the description of the processing analytical details applied to fMRI data are considered necessary for interpretation and replicability purposes, likewise QC procedures are instrumental to results interpretation. Thus, QC reporting should become an integral part of neuroimaging studies.

6. Conclusion

In this study, we presented the CONN quality control pipeline for the visual and automated QC testing of resting state fMRI data for FC-MRI analysis, demonstrated on publicly available and

heterogeneous data. We complemented knowledge and guidelines from the literature with additional automated QC strategies. Several, modular, and mutually non-exclusive procedures were included and emphasized how automated QC testing can help guide choices of preprocessing, denoising, and exclusion procedures. Overall, visual and automated QC were reciprocally informative, and their synergy was necessary for a sensitive evaluation of fMRI quality at all stages of the data life cycle. We hope this work contributes to the understanding, dissemination, and standardization of QC testing and QC reporting among peers and in scientific journals.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: fMRI Open QC Project, <https://osf.io/qaesm/files/osfstorage>.

Ethics statement

The studies involving human participants were reviewed and approved by fMRI Open QC Project. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

FM, AN-C, and SW-G contributed to the design of the study, data analysis, data interpretation, and manuscript writing. AN-C developed and maintains the CONN toolbox software. All authors revised and approved the submitted manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2023.1092125/full#supplementary-material>

References

- Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. *NeuroImage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034
- Andersson, J. L. R., Hutton, C., Ashburner, J., Turner, R., and Friston, K. (2001). Modeling geometric deformations in EPI time series. *NeuroImage* 13, 903–919. doi: 10.1006/nimg.2001.0746
- Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *NeuroImage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018
- Backhausen, L. L., Herting, M. M., Buse, J., Roessner, V., Smolka, M. N., and Vetter, N. C. (2016). Quality control of structural MRI images applied using FreeSurfer—A hands-on workflow to rate motion artifacts. *Front. Neurosci.* 10:558. doi: 10.3389/fnins.2016.00558
- Behzadi, Y., Restom, K., Liu, J., and Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* 37, 90–101. doi: 10.1016/j.neuroimage.2007.04.042
- Benhajali, Y., Badhwar, A., Spiers, H., Urchs, S., Armoza, J., Ong, T., et al. (2020). A standardized protocol for efficient and reliable quality control of brain registration in functional MRI studies. *Front. Neuroinform.* 14:7. doi: 10.3389/fninf.2020.00007
- Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S. G., de Zwart, J. A., Shmueli, K., et al. (2009). Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: A 7 T study. *Magn. Reson. Imaging* 27, 1019–1029. doi: 10.1016/j.mri.2009.02.004
- Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 84–88. doi: 10.1038/s41586-020-2314-9
- Caballero-Gaudes, C., and Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage* 154, 128–149. doi: 10.1016/j.neuroimage.2016.12.018
- Calhoun, V. D., Wager, T. D., Krishnan, A., Rosch, K. S., Seymour, K. E., Nebel, M. B., et al. (2017). The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. *Hum. Brain Mapp.* 38, 5331–5342. doi: 10.1002/hbm.23737
- Chai, X. J., Castañón, A. N., Öngür, D., and Whitfield-Gabrieli, S. (2012). Anticorrelations in resting state networks without global signal regression. *NeuroImage* 59, 1420–1428. doi: 10.1016/j.neuroimage.2011.08.048
- Chou, Y., Chang, C., Remedios, S. W., Butman, J. A., Chan, L., and Pham, D. L. (2022). Automated classification of resting-state fMRI ICA components using a deep Siamese network. *Front. Neurosci.* 16:768634. doi: 10.3389/fnins.2022.768634
- Churchill, N. W., and Strother, S. C. (2013). PHYCAA+: An optimized, adaptive procedure for measuring and controlling physiological noise in BOLD fMRI. *NeuroImage* 82, 306–325. doi: 10.1016/j.neuroimage.2013.05.102
- Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage* 154, 174–187. doi: 10.1016/j.neuroimage.2017.03.020
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance Neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Craddock, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S. S., Yan, C., et al. (2013). Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC). *Front. Neuroinform. Conference Abstract: Neuroinformatics* 7:42. doi: 10.3389/conf.fninf.2013.09.00042
- De Blasi, B., Caciagli, L., Storti, S. F., Galovic, M., Koeppe, M., Menegaz, G., et al. (2020). Noise removal in resting-state and task fMRI: Functional connectivity and activation maps. *J. Neural Eng.* 17:046040. doi: 10.1088/1741-2552/aba5cc
- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2013). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- DuPre, E., Salo, T., Ahmed, Z., Bandettini, P., Bottenhorn, K., Caballero-Gaudes, C., et al. (2021). TE-dependent analysis of multi-echo fMRI with tedana. *J. Open Source Softw.* 6:3669. doi: 10.21105/joss.03669
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* 12:e0184661. doi: 10.1371/journal.pone.0184661
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* 16, 111–116. doi: 10.1038/s41592-018-0235-4
- Fischl, B. (2012). FreeSurfer. *NeuroImage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021
- Friedman, L., and Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *J. Magn. Reson. Imaging* 23, 827–839. doi: 10.1002/jmri.20583
- Friedman, L., and Glover, G. H. The FBIRN Consortium (2006). Reducing interscanner variability of activation in a multicenter fMRI study: Controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *NeuroImage* 33, 471–481. doi: 10.1016/j.neuroimage.2006.07.012
- Friston, K. J., and Al, E. (2007). *Statistical parametric mapping: The analysis of functional brain images*. London: Academic.
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. J., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312
- Glover, G. H., Mueller, B. A., Turner, J. A., van Erp, T. G. M., Liu, T. T., Greve, D. N., et al. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *J. Magnet. Reson. Imaging* 36, 39–54. doi: 10.1002/jmri.23572
- Golestani, A. M., and Chen, J. J. (2022). Performance of temporal and spatial independent component analysis in identifying and removing low-frequency physiological and motion effects in resting-state fMRI. *Front. Neurosci.* 16:867243. doi: 10.3389/fnins.2022.867243
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. doi: 10.1038/sdata.2016.44
- Greve, D. N., Brown, G. G., Mueller, B. A., Glover, G., and Liu, T. T. (2012). A survey of the sources of noise in fMRI. *Psychometrika* 78, 396–416. doi: 10.1007/s11336-012-9294-0
- Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M. F., et al. (2017). Hand classification of fMRI ICA noise components. *NeuroImage* 154, 188–205. doi: 10.1016/j.neuroimage.2016.12.036
- Hagler, D. J., Hatton, S. N., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., et al. (2019). Image processing and analysis methods for the adolescent brain cognitive development study. *NeuroImage* 202:116091. doi: 10.1016/j.neuroimage.2019.116091
- Hallquist, M. N., Hwang, K., and Luna, B. (2013). The nuisance of nuisance regression: Spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. *NeuroImage* 82, 208–225. doi: 10.1016/j.neuroimage.2013.05.116
- Henson, R. N. A., Buechel, C., Josephs, O., and Friston, K. J. (1999). The slice- timing problem in event-related fMRI. *NeuroImage* 9, 1–125.
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17, 825–841. doi: 10.1006/nimg.2002.1132
- Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* 264, 47–56. doi: 10.1016/j.jneumeth.2016.03.001
- Liu, T. T. (2016). Noise contributions to the fMRI signal: An overview. *NeuroImage* 143, 141–151. doi: 10.1016/j.neuroimage.2016.09.008
- Liu, T. T., and Falahpour, M. (2020). Vigilance effects in resting-state fMRI. *Front. Neurosci.* 14:321. doi: 10.3389/fnins.2020.00321
- Liu, T. T., Glover, G. H., Mueller, B. A., Greve, D. N., Rasmussen, J., Voyvodic, J. T., et al. (2015). “Quality assurance in functional MRI,” in *fMRI: From nuclear spins to brain functions. Biological magnetic resonance*. eds. K. Uludag, K. Ugurbil and L. Berliner (Boston, MA: Springer).
- Lu, W., Dong, K., Cui, D., Jiao, Q., and Qiu, J. (2019). Quality assurance of human functional magnetic resonance imaging: A literature review. *Quant. Imaging Med. Surgery* 9, 1147–1162. doi: 10.21037/qims.2019.04.18
- Maknojia, S., Churchill, N. W., Schweizer, T. A., and Graham, S. J. (2019). Resting state fMRI: Going through the motions. *Front. Neurosci.* 13:825. doi: 10.3389/fnins.2019.00825
- Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., et al. (2013). Human connectome project informatics: Quality control, database services, and data visualization. *NeuroImage* 80, 202–219. doi: 10.1016/j.neuroimage.2013.05.077
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The open neuro resource for sharing of neuroscience data. *eLife* 10, 1–17. doi: 10.7554/eLife.71774
- Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage* 44, 893–905. doi: 10.1016/j.neuroimage.2008.09.036
- Nieto-Castanon, A. (2020). *Handbook of functional connectivity magnetic resonance imaging methods in CONN*. Boston, MA: Hilbert Press. doi: 10.56441/hilbertpress.2207.6598
- Nieto-Castanon, A. (2022). Preparing fMRI data for statistical analysis. arxiv: 2210.13564 [q-bio]. Available at: <https://arxiv.org/abs/2210.13564> (Accessed November 8, 2022).

- Nieto-Castanon, A., and Whitfield-Gabrieli, S. (2022). *CONN functional connectivity toolbox: RRID SCR_009550, release 22*. Boston, MA: Hilbert Press. doi: 10.56441/hilbertpress.2246.5840
- Parkes, L., Fulcher, B., Yücel, M., and Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage* 171, 415–436. doi: 10.1016/j.neuroimage.2017.12.073
- Power, J. D. (2017). A simple but useful way to assess fMRI scan qualities. *NeuroImage* 154, 150–158. doi: 10.1016/j.neuroimage.2016.08.009
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* 84, 320–341. doi: 10.1016/j.neuroimage.2013.08.048
- Raamana, P. R., Theyers, A., Selliah, T., Bhati, P., Arnott, S. R., Hassel, S., et al. (2020). Visual QC protocol for FreeSurfer cortical Parcellations from anatomical MRI. *bioRxiv*. doi: 10.1101/2020.09.07.286807
- Saad, Z. S., Reynolds, R. C., Jo, H. J., Gotts, S. J., Chen, G., Martin, A., et al. (2013). Correcting brain-wide correlation differences in resting-state FMRI. *Brain Connect.* 3, 339–352. doi: 10.1089/brain.2013.0156
- Sikka, S., Cheung, B., Khanuja, R., Ghosh, S., Yan, C., Li, Q., Vogelstein, J., Burns, R., Colcombe, S., Craddock, C., Mennes, M., Kelly, C., Dimartino, A., Castellanos, F. and Milham, M. (2014). Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC). The Configurable Pipeline for the Analysis of Connectomes (C-PAC). 5th INCF Congress of Neuroinformatics, Munich, Germany.
- Storelli, L., Rocca, M. A., Pantano, P., Pagani, E., De Stefano, N., Tedeschi, G., et al. (2019). MRI quality control for the Italian neuroimaging network initiative: Moving towards big data in multiple sclerosis. *J. Neurol.* 266, 2848–2858. doi: 10.1007/s00415-019-09509-4
- Strother, S., La Conte, S., Kai Hansen, L., Anderson, J., Zhang, J., Pulapura, S., et al. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I A preliminary group analysis. *NeuroImage* 23, S196–S207. doi: 10.1016/j.neuroimage.2004.07.022
- Taylor, P., Etzel, J., Glen, D., Reynolds, R., Moraczewski, D., and Basavaraj, A. (2022). FMRI open QC project. Available at: <https://osf.io/qaesm/>
- Tong, Y., Hocke, L. M., and Frederick, B. B. (2019). Low frequency systemic hemodynamic 'noise' in resting state BOLD fMRI: Characteristics, causes, implications, mitigation strategies, and applications. *Front. Neurosci.* 13:787. doi: 10.3389/fnins.2019.00787
- Wang, P., Wang, J., Michael, A., Wang, Z., Klugah-Brown, B., Meng, C., et al. (2021). White matter functional connectivity in resting-state fMRI: Robustness, reliability, and relationships to gray matter. *Cereb. Cortex* 32, 1547–1559. doi: 10.1093/cercor/bhab181
- Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* 2, 125–141. doi: 10.1089/brain.2012.0073
- Whitfield-Gabrieli, S., Nieto-Castanon, A., and Ghosh, S. (2011). *Artifact detection tools (ART), Release version 7.11*. Cambridge, MA: Artifact Detection Tools.
- Williams, B., and Lindner, M. (2020). PyfMRIqc: A software package for raw fMRI data quality assurance. *J. Open Res. Softw.* 8:23. doi: 10.5334/jors.280
- Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., et al. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum. Brain Mapp.* 39, 4213–4227. doi: 10.1002/hbm.24241



OPEN ACCESS

EDITED BY

Jo Etzel,
Washington University in St. Louis,
United States

REVIEWED BY

Sungho Tak,
Korea Basic Science Institute (KBSI),
Republic of Korea
Changwei Wu,
Taipei Medical University,
Taiwan

*CORRESPONDENCE

Daniel A. Handwerker
✉ handwerkerd@nih.gov

SPECIALTY SECTION

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

RECEIVED 16 November 2022

ACCEPTED 17 March 2023

PUBLISHED 06 April 2023

CITATION

Teves JB, Gonzalez-Castillo J, Holness M,
Spurney M, Bandettini PA and
Handwerker DA (2023) The art and science of
using quality control to understand and
improve fMRI data.
Front. Neurosci. 17:1100544.
doi: 10.3389/fnins.2023.1100544

COPYRIGHT

© 2023 Teves, Gonzalez-Castillo, Holness,
Spurney, Bandettini and Handwerker. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

The art and science of using quality control to understand and improve fMRI data

Joshua B. Teves¹, Javier Gonzalez-Castillo¹, Micah Holness¹,
Megan Spurney¹, Peter A. Bandettini^{1,2} and
Daniel A. Handwerker^{1*}

¹Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States, ²Functional MRI Core Facility, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States

Designing and executing a good quality control (QC) process is vital to robust and reproducible science and is often taught through hands on training. As FMRI research trends toward studies with larger sample sizes and highly automated processing pipelines, the people who analyze data are often distinct from those who collect and preprocess the data. While there are good reasons for this trend, it also means that important information about how data were acquired, and their quality, may be missed by those working at later stages of these workflows. Similarly, an abundance of publicly available datasets, where people (not always correctly) assume others already validated data quality, makes it easier for trainees to advance in the field without learning how to identify problematic data. This manuscript is designed as an introduction for researchers who are already familiar with fMRI, but who did not get hands on QC training or who want to think more deeply about QC. This could be someone who has analyzed fMRI data but is planning to personally acquire data for the first time, or someone who regularly uses openly shared data and wants to learn how to better assess data quality. We describe why good QC processes are important, explain key priorities and steps for fMRI QC, and as part of the FMRI Open QC Project, we demonstrate some of these steps by using AFNI software and AFNI's QC reports on an openly shared dataset. A good QC process is context dependent and should address whether data have the potential to answer a scientific question, whether any variation in the data has the potential to skew or hide key results, and whether any problems can potentially be addressed through changes in acquisition or data processing. Automated metrics are essential and can often highlight a possible problem, but human interpretation at every stage of a study is vital for understanding causes and potential solutions.

KEYWORDS

fMRI, quality control, neuroimaging, reproducibility, resting state, GLM, noise removal

1. Introduction

The fundamental question that a quality control (QC) process should answer is, “Will these data have the potential to accurately and effectively answer my scientific question and future questions others might ask with these data?” The secondary goal of QC is to identify data anomalies or unexpected variations that might skew or hide key results so that this variation can either be reduced through data processing or excluded. Even for a perfectly designed study,

problems can arise during nearly every step of the data acquisition and analysis. While a specific problem might be unexpected, the existence of problems should be expected. Failure to check the quality of data will result in incorrect or misleading interpretations of data. Therefore, a QC process should be a fundamental element in the design of any study. While good QC processes will not guarantee good results, they can greatly reduce the chances of generating misleading or incorrect results.

QC is both a key part of scientific progress in fMRI and a neglected topic. Overviews of good practices mention the importance of a good QC process (Poldrack et al., 2008; Nichols et al., 2017), but do not describe the elements of a good QC process in depth. Detailed QC protocols for fMRI studies tend to be published only for large or multi-site studies, do not always present context, and only a few include operating procedures for non-automated steps (Friedman and Glover, 2006; Marcus et al., 2013; Alfaro-Almagro et al., 2018; Kim et al., 2019; Scott et al., 2020; Huber et al., 2021; Huguet et al., 2021). Publications and seminars that systematically discuss and debate expectations and methods of QC for fMRI are rare. Automated or semi-automated QC tools have long been part of fMRI processing pipelines (Cox, 1996) and there is a growth in QC tools for specific phases of acquisition and processing (Dosenbach et al., 2017; Esteban et al., 2017; Heunis et al., 2020). Still, despite the central importance of good quality data for scientific reproducibility, there is only a modest amount of education and methods development research that focuses on improving QC processes.

Our anecdotal experience is that learning how to think about fMRI QC and the practical parts of checking data are often taught through hands-on training, particularly when people acquire data. With a rising number of researchers working with shared data and not acquiring data, a smaller proportion of neuroimagers may be receiving this necessary training during formative career stages. This is paired with an assumption that data that are published and shared are reasonable quality data. We have repeatedly heard shared datasets being referred to with terms such as “gold standard data,” which is another way of saying data users think they can trust downloaded data without running their own QC process.

To reduce these training gaps and push for more work and innovation, we document our approach to fMRI QC with two goals in mind: (1) Outline a quality control framework for fMRI for scientists who have not learned these skills during formative training periods. (2) Highlight QC priorities for a researcher who uses data they did not collect. We demonstrate a QC process, primarily using AFNI software, on a sample dataset as part of the FMRI Open QC Project.¹ For this project, multiple groups demonstrate their QC procedures with a variety of software packages on the same data.

While no manuscript can replace hands-on training, we highlight ways of thinking about fMRI QC that may guide additional learning. Our framework and demonstration are centered on the idea that automation should augment rather than replace human judgement. Also, discussions about QC often focus on what data to accept vs. exclude, but timely human judgement can identify problems that can be corrected through changes in

acquisition and analysis. This interaction between automation and human judgement will become more critical to understand and improve as fMRI datasets increase in size. Large studies require a clear plan for which aspects of QC can be automated and where the finite amount of human intervention and judgement is most useful. To that end, we provide a framework for thinking about general approaches with a specific focus on where human intervention is particularly important.

2. Quality control framework for fMRI

QC asks whether and how data can be used. For fMRI data, this comes down to addressing two questions (1) Which voxels have useable data? (2) Are the locations of those voxels in the brain accurately defined? Answers to the first question involve ensuring consistent fields of view across all scans, computing basic QC metrics such as signal-to-noise ratio (SNR) and the temporal-signal-to-noise ratio (TSNR), and searching for spatial and temporal artifacts which may render these areas unreliable for modeling. Answers to the second question involve looking at functional alignment between runs, functional to anatomical alignment, anatomical alignments to a common stereotaxic space, and anatomical alignments across study participants.

The quality checks needed to answer these questions are not the same for all study purposes and the best tools to answer them vary by study phase and purpose. As discussed in a generalized QC framework by (Wang and Strong, 1996), QC includes both intrinsic and contextual measures. Intrinsic measures characterize inherent properties of the data. For example, the average temporal-signal-to-noise ratio (TSNR) of gray matter voxels might be intrinsically useful. However, contextual measures depend upon the research hypothesis. For example, the TSNR values of voxels in the temporal pole might only matter in the context of studies with hypotheses about the temporal pole. Similarly, some functional-to-anatomical alignments are intrinsically poor, but an imperfect alignment might be sufficient in the context of a study that focuses on large regions-of-interest (ROIs) or spatially smoothed data. As another example, a modest amount of head motion or breathing artifacts might be addressable through data processing for some studies but could be problematic in the context of a study with task-correlated breathing (Birn et al., 2009) or with population biases in head motion (Power et al., 2012). This distinction between intrinsic and contextual quality is critical because many discussions of fMRI QC focus on whether to keep or exclude data, yet there are often situations where data can be processed to be useful for a subset of potential applications, underscoring the need to keep the application of data central when assessing quality.

We organize our QC framework into four phases: during study planning, during data acquisition, soon after acquisition, and during processing. This structure should guide when to think about certain steps, but the same overall issues cross all phases, and they are not in a strict temporal order. For example, an issue identified during processing may prompt changes to study design or acquisition. An additional element of QC is QC of the acquisition hardware, which should be checked regularly as part of the operational procedures of any fMRI research facility. Since there are already multiple resources for this type of fMRI QC (Friedman et al., 2006; Liu et al., 2015; Cheng

¹ QC Project main page: <https://www.frontiersin.org/research-topics/33922/demonstrating-quality-control-qc-procedures-in-fmri>

and Halchenko, 2020), we are limiting our scope to QC that is specific to the data collected during a study. The appendix summarizes the suggestions in this framework for use as a guide when designing a study-specific QC protocol.

2.1. QC during study planning

Good QC procedures depend on having the QC-relevant information stored in a representationally consistent manner where they can be efficiently accessed (Wang and Strong, 1996). This requires effort during the planning stage of a study to make sure this information will be identified, collected, and organized. Defining QC priorities during the planning phase also supports future data sharing. The information that needs to be organized to support a robust QC protocol will also be accessible to future users of the data.

Expert study-specific advice is highly recommended during study planning. If one has access to experts in experimental design and acquisition, seek out their advice during this phase rather than the “What is wrong with my data?” phase. Many of the QC protocols referenced in the introduction feature study-specific examples and show how others have prioritized and organized QC-relevant information. Key topics to consider when planning a study are:

- What QC measures will support the goals of the study? For example, if a study has *a priori* ROIs then QC measures for those ROIs and pilot scans that optimize those QC measures can flag issues that prompt acquisition changes and avoid wasted data.
- Minimize variability in operating procedures across scan sessions by generating checklists and written instructions that clearly describe what experimenters should do during the scan (e.g., acquisition instructions), and should tell to participants [e.g., clear task or rest instructions and protocols to decrease head motion (Greene et al., 2018)]. The same applies to preprocessing and QC measures to calculate soon after each scan so that issues can be efficiently identified. (Strand, 2023) is a general overview for how good procedures can help avoid errors and improve data quality.
- What data should be collected during acquisition that will support QC later? This includes both logs of expected and unexpected events such as: participant behavior (e.g., task behavioral response logs, feedback from participants, observed movement during runs, seemed to fall asleep in a run, needed to leave scanner & get back in), issues with stimulus presentation, qualitative observations and quantitative measures of real-time data quality, respiratory and cardiac traces, external sources of variation between participants [e.g., time of day, caffeine intake, endogenous and exogenous sex hormone variation (Taylor et al., 2020)] and all scanning parameters.
- How QC measures will be organized and shared. Acquisition-stage QC is useful only if it is connected to the data, understandable by others, and easy to share.
- Finally, pilot sessions should go beyond attempting to optimize MRI acquisition parameters, to play a role in addressing all the above QC topics, so that when acquisition for a study begins, the procedures for acquiring, organizing, and rapidly checking QC metrics are already in place.

2.2. QC during data acquisition

It is better to design and follow a QC-focused scanning protocol and proactively collect good data than to retrospectively attempt to remove or fix bad data. That means one should aim to look at reconstructed MRI data as soon as feasible to identify unusual dropout or serious artifacts. When scanners are equipped with real-time fMRI capabilities, this initial inspection can happen as volumes are being acquired. While all modern scanners allow people to look at volumes during a scanner session, additional, real-time systems such as AFNI (Cox and Jesmanowicz, 1999) and NOUS (Dosenbach et al., 2017) can help identify artifacts in time series and excessive motion events, prompting researchers to notify the participant and to re-collect data. Real-time quality checks should be extended to any concurrent peripheral measurements such as respiratory or cardiac traces, behavioral responses, EEG, and eye tracking, to name a few. Stimulus presentation scripts can also integrate some rapid feedback so that experimenters can identify participants who are not performing a task as expected. Even if a session-specific issue observed during acquisition is not correctable in real-time, it can be flagged during acquisition for closer attention during processing or can lead to protocol changes to improve future scanning sessions.

2.3. QC soon after acquisition or download

Rapid QC after acquisition can focus on intrinsic issues that might not have been obvious during acquisition. If done between acquisition sessions, information gathered this way can identify ways to improve future acquisitions and avoid unexpected downstream analysis problems. The most important thing to check is that the expected data are present, have understandable and accurate file names, and are properly documented. Shared datasets often have a few surprises (e.g., missing or corrupted files, duplicated data, incomplete runs). For example, early QC can help identify and fix a task presentation script that insufficiently logged behavioral responses and times. These early checks should also include confirming that each MRI run and peripheral measurement, such as respiration and cardiac traces, have the correct number of samples, and look as expected. Checks should also determine if fMRI data look anatomically correct and have consistent orientation and brain coverage. This should also include checking whether parameters in data headers are plausible and match documentation. For example, we recently saw a dataset where the publication accurately listed a slow 5.1 s TR for a specialized sequence, but the files were incorrectly saved with a 1.5 s TR in their headers. This caused problems when processing steps read the incorrect TR from the file headers.

For shared data, check if there is any information about the QC procedure or a list of excluded runs or participants. If there is no information on problems with the data, that is likely a warning sign that there was no systematic QC procedure, and one should examine the data more carefully before using. If there was a clear QC procedure, one can also check if contextual metrics for newly planned analyses were included. For example, if the initial analysis focused on task responses and new plans focus on connectivity measures, the initial QC may not have focused on potential temporally correlated artifacts.

While full processing of data can be a slow process, an initial, limited preprocessing aimed at generating key automated QC metrics should be run as soon as possible. Even if a full

preprocessing pipeline is not finalized, running some basic preprocessing steps can identify issues that will help tally what data are useable and can help better optimize the final preprocessing pipeline. For example, if anatomical to functional alignment is poor in many participants during initial preprocessing, then time can be devoted to figuring out ways to optimize the alignment algorithms for a given dataset.

2.4. QC during data processing

The big advantage of integrating QC into a data processing pipeline is that QC metrics and key images for visual inspection can be automatically calculated for multiple steps in the pipeline. For example, AFNI's `afni_proc.py` pipeline automatically generates a QC html page with values and images that aid human interpretation of data quality. By compiling automatically calculated measures, someone with modest training can view reports to identify many things that look odd and are worth showing to a more experienced researcher.

While the processing steps have a fixed order, examination and interpretation of QC measures do not. Therefore, automated QC pipelines should calculate and organize measures from across the processing stream to aid human interpretation. This is particularly true for shared data where issues with unprocessed data may not have been checked or documented. For example, a few authors were recently working with a shared dataset where the acquired slices did not cover the most superior 5 mm of the cortex. This was flagged as a failure of the registration algorithm, but by going back to the unprocessed data, it became clear that the alignment was fine, but data were missing.

After data are processed, check if there are any warnings or errors from the execution of the processing script. These may seem obvious, but subtle downstream errors from unnoticed script failures happen. This is also the easiest place to see if the same warnings repeatedly appear and warrant changes to a processing pipeline. AFNI makes this easy by compiling the warnings from all processing steps in AFNI's QC output so that users can look in one place to see if any parts of the script failed to execute or if serious data issues were automatically flagged.

Then quality checks can be separated into answering the two questions from the beginning of this section: (1) Which voxels in a dataset have usable data? (2) Are the locations of those voxels in the brain accurately defined?

2.4.1. QC during data processing: Usable voxels

The most straightforward check is noting areas of the brain that were included in the scan's field-of-view. Since most pipelines attempt to mask out non-brain voxels, one must make sure the mask is not excluding brain voxels or retaining voxels outside the brain. fMRI data always suffers from signal dropout and distortions, so voxels within the brain are expected to be missing, but, for a study with the same acquisition parameters, the location and amount of dropout and distortion should be relatively consistent. A dataset with unusually large amounts of dropout should be checked to see if there are other issues. Even if dropout is fairly consistent, the QC process should identify voxels with usable data in only a subset of participants. Particularly for ROI-based analyses and connectivity measures, voxels with data in only a fraction of a population can cause non-trivial biases in data that are hidden under ROI averages or averaged group maps.

The temporal signal-to-noise ratio (TSNR = detrended mean/standard deviation) is a rough, but useful measure of fMRI quality that highlights issues that can be missed by looking only at the magnitudes, since the standard deviation of time series will be affected by temporal acquisition artifacts and head motion spikes. On a voxel-wise map, the spatial pattern of TSNR values can vary based on acquisition options. For example, a 64-channel head coil with many small receiver coils will likely have relatively higher TSNR values on the surface versus the middle of the brain compared to a 16-channel coil (although the raw TSNR values should be higher everywhere). In addition to viewing TSNR maps, with consistent acquisition parameters, TSNR should be similar across a study, so data warrants closer examination if the average TSNR for the whole brain, white matter, or gray matter is lower in some runs.

Mean images and TSNR are useful for identifying potential problems, but not necessary for understanding causes and potential solutions. By recognizing different types of MRI artifacts, it is possible to figure out if a problem can be solved through data processing, or censoring time points or voxels. Not every artifact is a problem. For example, the differences in TSNR between the surface & the center of the brain with multi-channel head coils is not inherently a problem, but it can affect studies that directly compare or correlate cortical surface and subcortical responses (Caparelli et al., 2019). MRI imaging artifacts are best understood with hands-on training, but there are some key things to look for. Any contrast changes that do not seem to follow brain tissue or are not symmetric between hemispheres might be artifacts. It is important to look at data from multiple views (i.e., axial and sagittal) because some artifacts may be obvious within acquired slices and others may be visible across slices. If there is a bright artifact in one location, it might be possible to exclude data from that location, but many types of artifacts are obvious in one location and present, but less obvious over a larger portion of the brain, which would make data unusable. Processing that includes masking or temporal scaling of the data can often hide these artifacts, but they can be more visible in TSNR versus mean images or if the contrast is adjusted to give values nearer to zero more brightness. Another useful tool is to look at power spectra of data, which can identify if an artifact is fluctuating at consistent frequencies. Temporally periodic artifacts can be due to acquisition problems that might affect an entire dataset or by respiratory and cardiac fluctuations which are potentially addressable.

If the brain volume overlaps itself or there is a replicated part of the brain where it should not be, this wrapping or ghosting can inject signal from one part of the brain into another part and make a run unusable. A way to examine the seriousness of a ghosting or wrapping artifact is to correlate the rest of the brain to voxels within the artifact. AFNI's *instacorr* interface lets users interactively correlate data to specified voxels and is particularly useful for this. *Instacorr* does not depend on AFNI processing so it can be used on data processed with other packages. If a voxel in an artifact is correlated with other clusters of voxels in a non-anatomical pattern (e.g., The signal in one brain region correlates with the same-shaped ghosted region elsewhere in the volume) that is a serious sign that the artifact corrupted the data.

One additional tool for identifying temporal artifacts in voxels is to look at partially-thresholded and unmasked activation maps for both task-locked GLM models and correlations to the global averaged signal or white matter. While one cannot reject a dataset if the task of interest is not significant, if a study uses a visual task and there is no

task-locked activity in the primary visual cortex, then there are likely additional issues with the data. If there is task-locked activity outside of the brain or on tissue/CSF boundaries, that is a sign of ghosting, motion artifacts, or task-locked breathing (Birn et al., 2009). If there is not a task, correlation maps can highlight similar issues, but they can also be used to identify population differences. For example, given the widely documented differences in global signal across populations (Power et al., 2012; Gotts et al., 2013; Yang et al., 2014), any study that plans to regress out the global signal as noise needs to correlate the global signal to the other voxels in the brain and test whether the correlation between the global signal and voxels systematically varies between populations or other contrasts of interest.

There are many automated QC metrics, in addition to TSNR, that can be used to automatically exclude data in voxels or highlight areas of concern. The most common ones are spike detection and motion estimates. Those can be used to both censor specific volumes and to automatically decide whether a run has too many censored volumes to be useable. The remaining degrees of freedom (DOF) after temporal filtering, censoring, and noise regression can be used to decide if sufficient DOF remain for statistical tests. The effect of temporal filtering on the loss of degrees of freedom is sometimes ignored in fMRI studies. AFNI also outputs a spatial smoothness estimate for each dataset. These numbers are not especially useful in a single run, but for a given set of acquisition parameters, the smoothness estimate should be roughly consistent across a study. If smoothness estimates vary widely, it is worth looking more carefully at outlier runs.

2.4.2. QC during data processing: Alignment

Evaluating individual voxel data quality benefits greatly from automation, but masking and alignment results often require manual inspection and interpretation. This is because different acquisitions can have different contrasts and parameters, so what works well for one dataset might not work as well for another. Artifacts and non-trivial spatial distortions in unprocessed data can also affect masking and alignment. Automated metrics for alignment quality will keep improving, such as with a metric to automatically warn that the left and right sides of the brain are flipped (Glen et al., 2020). Automation can be used to compile images that facilitate human inspection. AFNI's html reports include images where the sulcal edges from a participant's anatomical volume are overlayed onto the functional images or common anatomical templates. This is a quick way to catch clearly mis-aligned brain edges or sulci and potential issues that are worth a closer examination of the full volumes' alignments.

Visual checks can focus on several factors. If collected during the same session, an anatomical image should have a decent alignment to the functional data even without processing. Atypical brain structures can be viewed before processing. An expert can tell which types of variation are concerning – either to the volunteer or to data processing – but a less experienced reviewer can flag anything that is asymmetric for expert review. Benign cysts, larger ventricles, and other atypical structures do not require rejecting data, but they can affect spatial alignment between participants as well as the locations of functional brain areas. As such, those occurrences should be noted, and more attention should be spent on assessing alignment quality.

Since most fMRI research uses multi-channel receiver coils, one very common artifact is intensity inhomogeneity, where the voxels closest to the head coil have a higher magnitude signal than voxels

nearer to the center of the brain. This inhomogeneity can look bad, but it is not inherently a problem. That said, it can affect the accuracy of brain masking and alignment so, if the data has a lot of inhomogeneity, it is useful to spend more time checking brain masking and alignment.

It is worth taking time to make sure a brain mask excludes sinuses and non-brain tissue, and that a mask does not remove parts of the brain. Inconsistent masking often leads to flawed anatomical-to-functional alignment and flawed reregistration between participants. Unless problems are caused by artifacts or distortions, it is often possible to fix alignment issues by tuning function parameters or by hand-editing masks.

Once many participants in a study are processed and aligned to a template, a summation of all the fMRI coverage maps is very useful for identifying brain regions that are included in only a portion of study participants. Excessive blurring on the average of the aligned images can also signal faulty alignment for a subset of participants. From our experience, looking at such coverage maps is strangely uncommon. A concatenated time series of all anatomical images and an average anatomical are very useful for checking the consistency of alignment across a population.

2.5. Peripheral measures

QC for fMRI studies often focuses on the MRI data, but unprocessed and processed peripheral measurements can also be sources of error. While many peripheral measures can be collected and checked, we will highlight a few examples for how to think about such measures in general. To be used with fMRI, peripheral measures need to log their timing in relation to fMRI volume acquisitions. Errors can arise in peak detection for respiratory and cardiac traces. Movement of a finger within a pulse oximeter can create noisy sections with what looks like rapid changes in heart rate that can negatively affect some peak detection algorithms. Anyone who collects respiratory data will also find spontaneous breath holds, which will affect fMRI data. Breath holds will cause large, brain-wide signal changes that bias results or merely be a non-trivial source of noise. For task-based fMRI, check response logs to confirm the expected information was logged and participants were compliant with task demands. Also check to make sure that head motion or respiration patterns are not task-correlated, since non-neural signal sources that are task-locked will bias results.

For all QC steps, it is crucial to consider that algorithms often fail in subtle ways rather than with clear errors, and these are the hardest errors to catch. It is therefore imperative that all steps be thoroughly vetted to ensure all assumptions required by the program are met and that programs are used consistently with their documented intent.

3. Methods

The previous section contains information on how to think about planning QC for a dataset. The following examples on shared datasets show how some of these concepts work in practice. As already noted, a QC process checks both intrinsic quality measures and contextual measures that are often dependent on the scientific question that a researcher has in mind. Additionally, because the data have already

been collected, we do not demonstrate the phases of QC before and during data collection (though in the discussion we will note some operational steps that could have been taken with these data). Since, we do not know the intended purposes for these data, we can make some assumptions about context, but our attention will primarily focus on intrinsic QC. We are focusing our contextual QC on issues that might affect connectivity measures for rest data or task responses for task data, without making assumptions about regions of interest.

We classify data which we believe could answer such questions as “included,” data which could not answer common or basic questions as “excluded,” and data which may be suitable for some questions but not others as “unsure.” Automation scripts were used to ensure consistency across subjects; the full processing and figure generation code and instructions may be found in our GitHub repository.² Each processing step was given its own script with the expectation that users could check results before proceeding.

Data were initially checked using a basic visual inspection to identify anything of concern in the data including missing information, artifacts, whether the image field of view included the whole brain (excluding the cerebellum and brain stem), and whether there were noticeable anatomical or image abnormalities. Concerns were noted, and screenshots were uploaded to a shared folder. Anything requiring additional discussion prompted either a message or a video chat between researchers to either (a) decide that the object of concern was inconsequential or (b) properly identify the problem and mark it.

For processing of the data after these inspections, T1 anatomical images were segmented using freesurfer's *recon-all* (Fischl and Dale, 2000), and a non-linear transformation for warping anatomical images to the MNI template space was calculated using AFNI's *@SSwarper* (Cox, 1996). *SSwarper*'s output includes QC images, which were checked both to make sure that the brain mask had complete coverage and did not include skull, and that the individual brain had been properly aligned to the MNI template.

AFNI's *afni_proc.py* program was used to perform slice timing correction, rigid-body motion correction, alignment of anatomical and echo-planar images, blurring to 6 mm full-width half-maximum, and regression of physiological-and motion-related signals. Volumes which contained more than 0.25 mm of head motion from neighboring volumes were censored. Voxels which were determined to be outliers by AFNI's *3dToutcount* were tallied and volumes which had more than 5% of voxels as outliers were censored.

For all data, the ANATICOR method (Jo et al., 2010) was used to compute regressors associated with scanner instabilities and physiological noise. In addition, we also regressed motion estimates and their first derivatives. For rest data (subjects 101–120), additional regressors were used to bandpass between 0.01 and 0.1 Hz, which significantly reduced the remaining DOF for the data. For task data, this step was omitted.

In the case of task data (subjects 001–030), tasks were modeled using the simplified task timings supplied with the data. The labeled task conditions were “control” and “task,” and each trial had an onset time and duration. We modeled task responses in our GLM with

AFNI's default double-gamma hemodynamic response function using both the onset and duration information.

For inspecting the outputs of all other steps, we relied primarily on *afni_proc*'s webpage-based QC report. Many figures in this manuscript use QC images that were automatically compiled in this report. Automatic motion correction and outlier censoring were used to see whether subjects exceeded 20% of volumes censored; in these cases, subjects were excluded.

The echo planar image (EPI) to anatomical alignment was checked by ensuring that anatomical edges matched the gyral shapes on the EPIs, that the ventricles were aligned, and that the brain was not distorted to the point of being displaced past the anatomical boundary.

Anatomical-to-template alignment was checked by ensuring subject-warped edges matched the template image's edges, and, that the gyral shapes on both the anatomicals and the brain edges matched. The final EPI mask was checked to ensure it covered all likely areas of interest (i.e., those targeted by scientific inquiry).

Model fits for regressors of interest were examined to make sure that good fits were not spatially aligned with previously identified artifacts. A similar inspection was performed for seed-based correlation maps to make sure that the underlying correlation structure was free of artifactual patterns.

For the task data, while we do not know the expected patterns, the modeled task responses were examined to see if they presented a plausible design with a sufficient number of uncensored trials per task condition.

Lastly, the warnings automatically generated by *afni_proc* were checked: these include unusually high correlations with nuisance regressors, total percentage of censored volumes, pre-steady-state detection, possible left–right flips, and EPI variance line warnings. For likely left–right flips, without additional information, we cannot ascertain whether the EPI or anatomical has the correct orientation; thus, such subjects are marked for exclusion. EPI variance line warnings are a marker of potential temporal artifact and *instacorr* was used to examine potential artifacts for severity.

4. Results

The task data contained numerous problems during the initial visual inspection process. Across the dataset, dropout and distortion were substantial in the unprocessed images. There were also very visible motion artifacts (e.g., Figures 1, 2). Four subjects were all automatically excluded because more than 20% of volumes exceeded motion and outlier censoring thresholds. Most subjects showed substantial dropout in the temporal lobe and some showed cerebellar dropout (Figure 3A). Several subjects showed atypically high correlations between a white matter ROI and gray matter voxels and areas of highest activation to the full F test for the task outside of the brain or in CSF (Figure 4). Based on EPI variance line warnings, visual inspection with *instacorr* identified several subjects with non-trivial artifacts (Figure 5A). Additionally, multiple subjects showed mild to moderate correlations between the task and control condition timing, which reduces that statistical power to independently estimate effect sizes for the two conditions. Since we did not create the study design, we did not exclude any participants solely because of this correlation. In total, 14 subjects were marked for inclusion, 12 were marked for exclusion, and 4

² https://github.com/nimh-sfm/SFIM_Frontiers_Neuroimaging_QC_Project

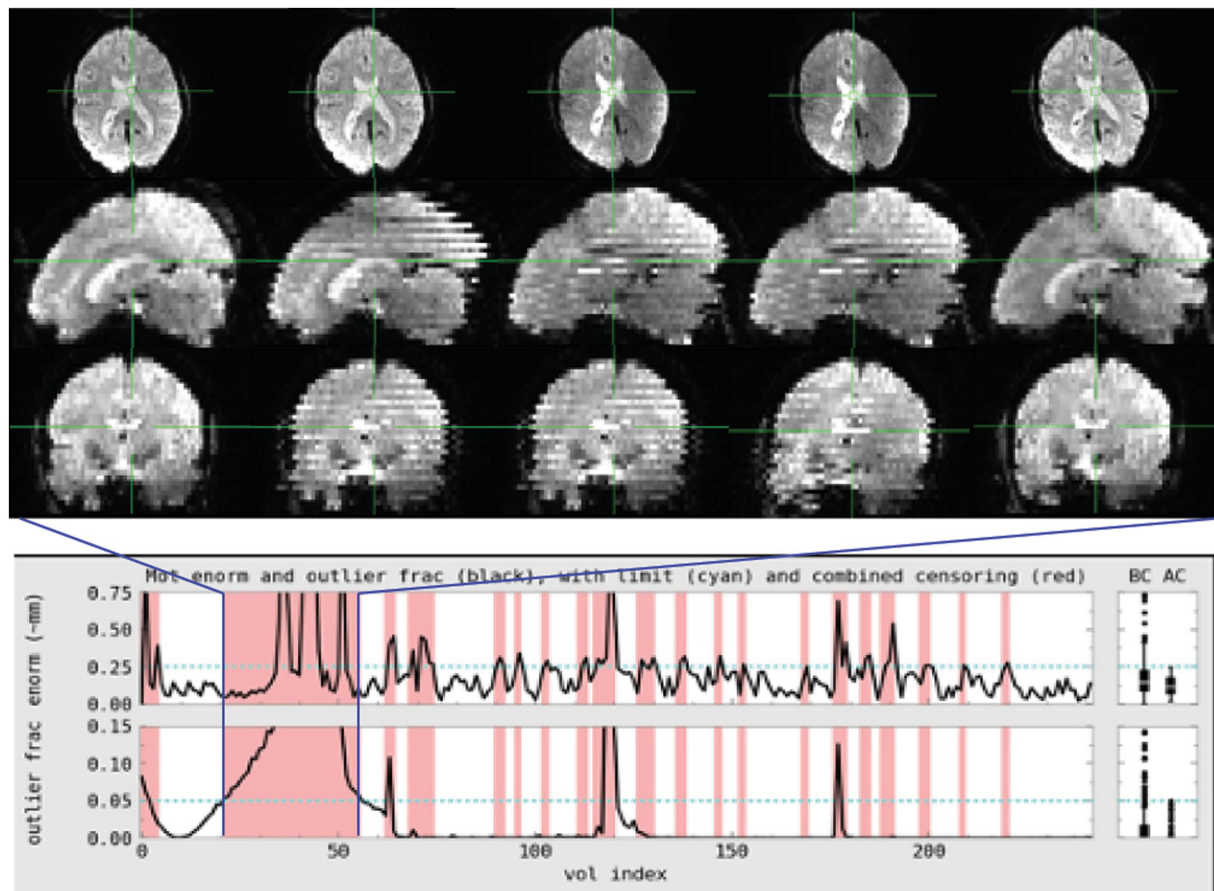


FIGURE 1

Subject 017 had a high number of censored volumes due to motion. This figure depicts several volumes in which the motion artifact is very clear. Banding due to the magnitude of head motion during acquisition are visible on the sagittal and coronal slices. Within the axial slice, this motion makes part of the lateral ventricles disappear because of displacement during acquisition. Such a large motion artifact should be visible on the console even in an axial-only view. Operationally, it would be useful to note this during acquisition and consider collecting an additional run while the subject is present.

were marked unsure out of a 30-subject data set. An overview of our findings across both datasets are shown in Table 1.

For the rest data, more of the brain was consistently covered (Figure 3B). Two subjects were automatically excluded because more than 20% of volumes exceeded motion and outlier censoring thresholds. Areas of general concern in the rest data included correlations between gray matter and a white matter ROI, poor correlations to expected networks from ROIs like the posterior cingulate, and EPI variance line warnings followed by *instacorr* inspection of artifacts. In these data, *instacorr* often showed issues related to EPI variance warnings in the unprocessed EPIs, but when censored volumes were removed by processing, *instacorr*-observable artifacts were reduced, and the remaining data were usable. The threshold between inclusion and exclusion based on these criteria was subjective, and the decision to exclude was typically based on several borderline reasons for concern, such as more than 10% of volumes censored and signs of artifacts in the data. We likely would have excluded more subjects if other subjects with this study were less noisy (Figure 6). Two subjects were excluded because the left and right sides of the brain were likely flipped between the anatomical and EPI data and an additional subject looked like the anatomical volume was from

a different brain than the EPI (Figures 7A–D). Given 3 participants showed an EPI and anatomical mismatch, there is a risk of an underlying issue with file naming and organization in these data. If we were using these data as part of a study, we would try to identify the origin of the flipping to confirm the scope of the problem and possibly identify the true left vs. right so that these participants would not need to be excluded. In total, 13 subjects were marked for inclusion and 7 for exclusion out of a 20-subject data set.

5. Discussion

We outlined priorities for QC of fMRI studies and then demonstrated them on two datasets. While priorities are best organized around conceptual goals, QC steps are ordered by when potentially serious problems are noticed. For the exemplar data, high motion, non-trivial distortion or dropout, and warnings signs for artifacts were rapidly apparent and dominated our focus. We highlight TSNR and several other measures as important QC metrics in our priorities, but we did not highlight them in practice. This is because some data did have low TSNR and artifacts that were

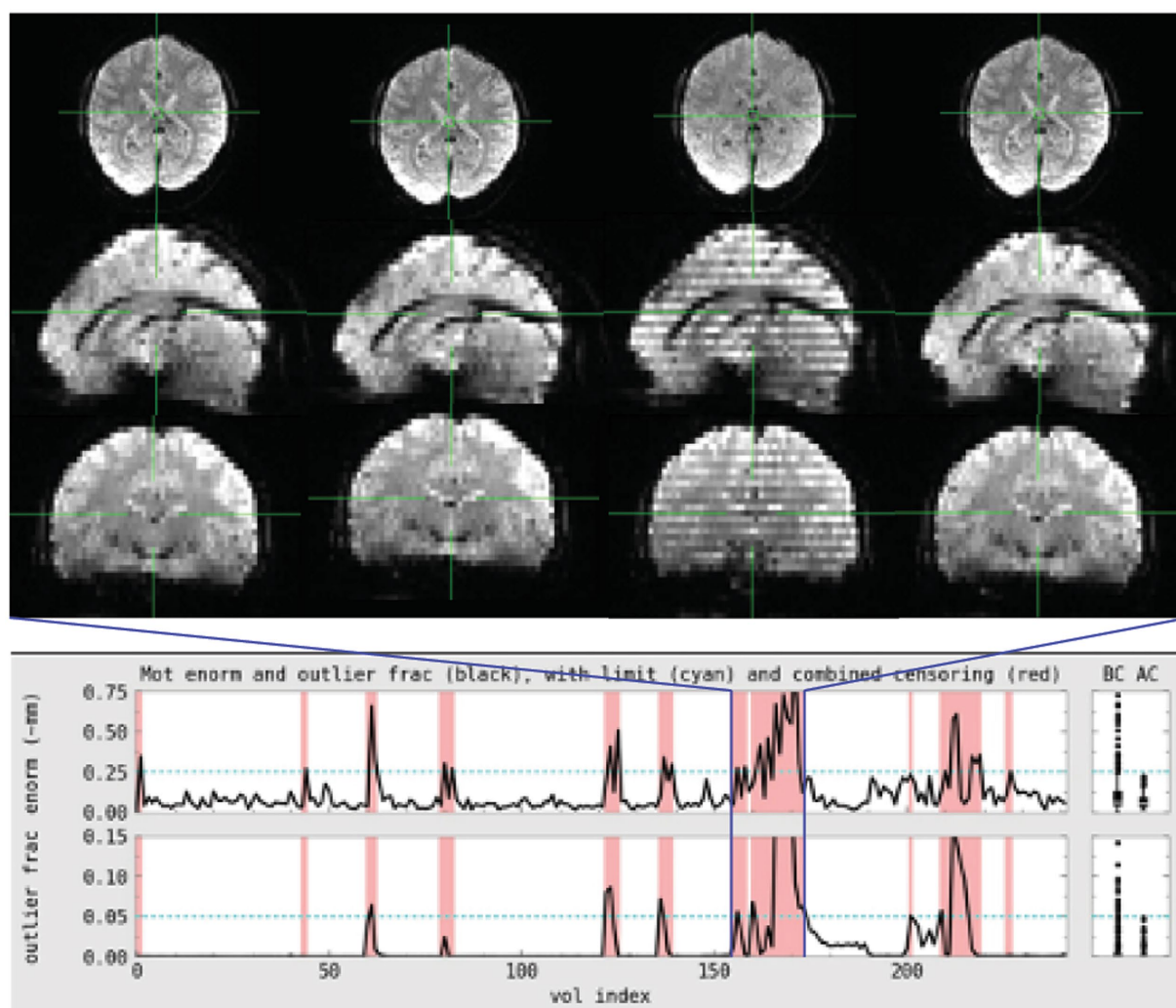


FIGURE 2

Subject 029 had a more subtle motion artifact than depicted for subject 017. The banding is visible during the period with the most motion but is otherwise more subtle and would be less likely to be noticed during acquisition without automated QC metrics.

clearly visible in TSNR maps, but these were in runs that were already rejected for other reasons. For these data, TSNR measures might have improved understanding of the effect of motion artifacts, but TSNR did not add value to decisions of what to include or exclude. In other datasets, TSNR has been the first place where something problematic is noticed.

This emphasizes a critical point of QC protocols. Datasets can have unique quirks, and the most useful QC checks for fMRI data are not universal across all studies. We've interacted with researchers who had a bad experience with head motion in a study and prioritized checks for head motion above all else. In fact, when the Organization for Human Brain Mapping put together a consensus statement on results reporting, it included a general recommendation to document QC measures, but only specified motion and incidental findings for fMRI data (Nichols et al., 2017). Reporting on alignment quality, MRI artifacts, degrees of freedom available, and consistency of the imaging field of view were not mentioned. For QC to become an intrinsic part of data acquisition, processing, and sharing, guidelines should be updated to include at least these valuable QC metrics.

A good QC process is designed to identify and address issues as soon as possible. The shared task data had many problems that were not addressable by the stage we received them. With the goal of improving the quality of shared data, we want to highlight QC steps that could have helped avoid collecting a dataset with such problems. Some problems, like the artifacts from extreme motion depicted in Figure 1, should have been observable during data acquisition. Real-time motion tracking, would identify high motion runs during scanning and potentially create an opportunity for additional acquisitions. Additional real-time monitoring of peripheral data, like eye tracking, behavioral responses, or cardiac and respiratory traces would identify drifts in consciousness or attention to the task. Once data are collected, rapidly running some subject-level analyses may identify correctable problems. For example, many of the acquisition issues in the rest data that might cause the spatio-temporal artifacts we saw would have been visible early in collection and might have been fixable through changes in acquisition. We reiterate that it is imperative to run analyses as early as possible to avoid acquiring large amounts of data with problems that do not arise until the study is analyzed months or years after acquisition began.

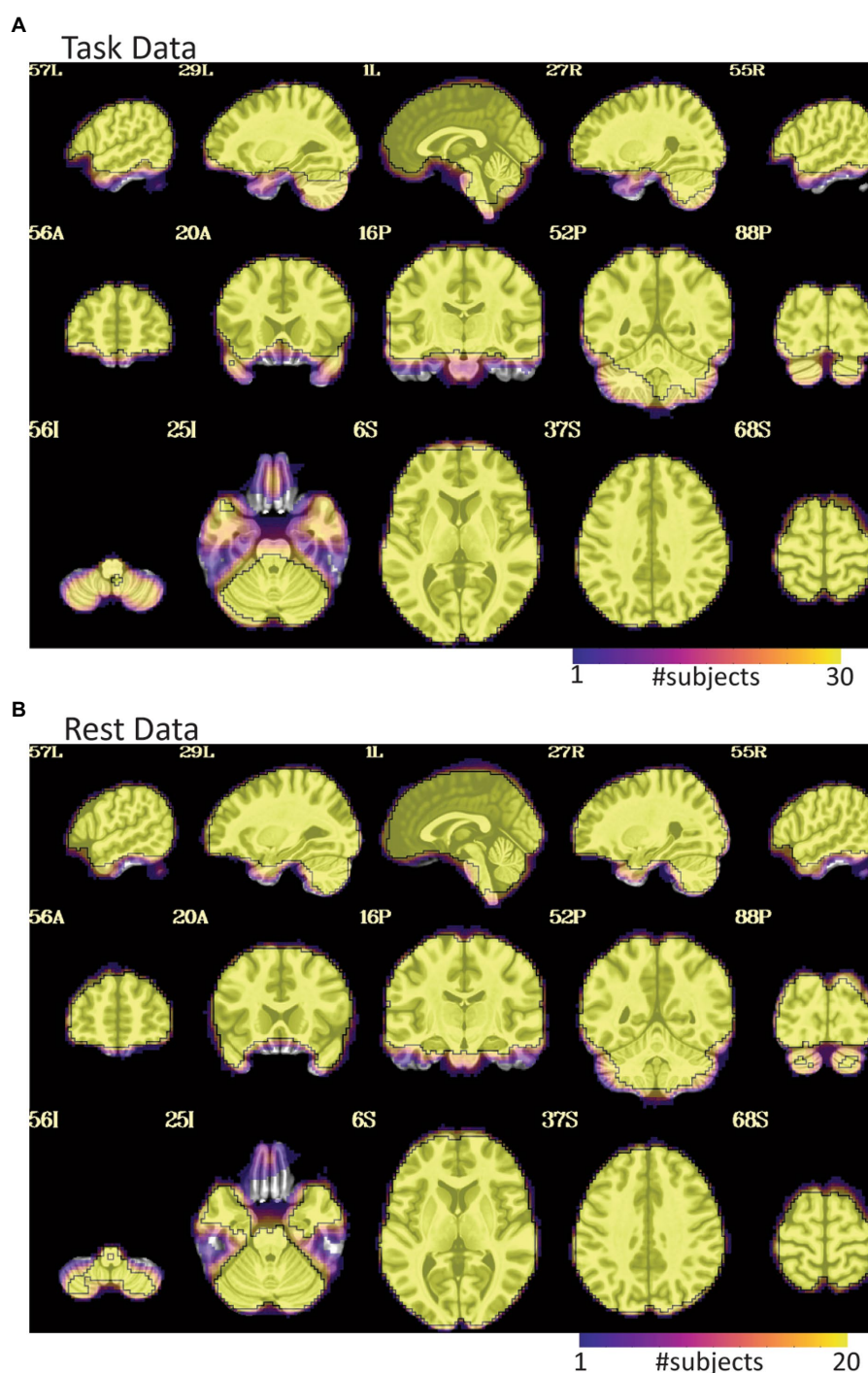


FIGURE 3

EPI coverage maps in MNI space for (A) task and (B) rest data sets. More yellow indicates that more subjects retained usable data for a given voxel. More purple indicates voxels where fewer subjects have usable data. The black outline surrounds voxels where all subjects have useable data. While both datasets show dropout in orbitofrontal and inferior temporal areas, the dropout is less consistent and more pervasive in the task data where much of the temporal lobe does not have usable data in a non-trivial fraction of subjects. The black line in (A) also highlights that not all subjects have cerebellar coverage.

Between the original and revised submission of this manuscript, we noticed a serious error in our processed data that we missed even while using a detailed QC protocol. For a subset of task subjects, the skull stripped anatomical volumes were mis-labeled and we aligned fMRI data to the wrong anatomicals. This created an unintentionally

good case-study on the limits of QC and how we could have caught this error earlier. We introduced this work by stating the purpose of QC is to identify whether data is of sufficient quality to be used for its intended purpose. In this case, we observed bad EPI to anatomical alignments, and wrote that the data would be not usable for their

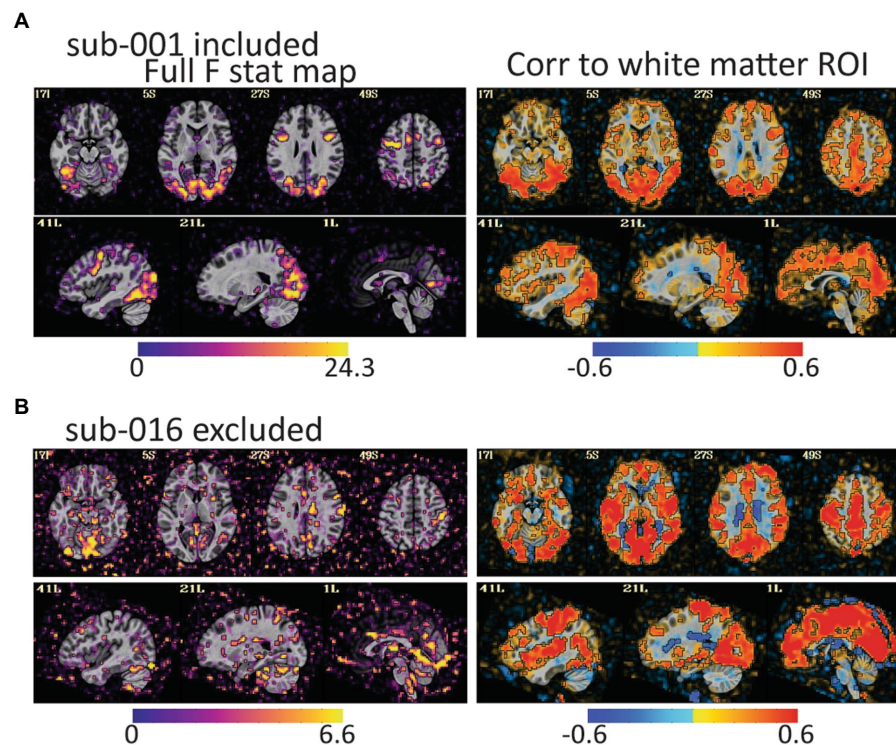


FIGURE 4

The full F stat map shows the decile of voxels with the highest F values for the full task GLM. The correlation to the white matter ROI shows voxels that correlated to white matter after the task design is regressed from the data. (A) For sub-001, F stat peaks are large and mostly in gray matter. (B) For sub-016, the F values are smaller, and the peaks are in lateral ventricles, CSF, and outside of the brain. The white matter correlation maps are harder to identify as clearly good or bad, but more pervasive correlations to gray matter as in (B) are an additional warning of a problem. Notably, both subjects have relatively little head motion (1.7% of volumes censored for sub-001 and 3.7% of volumes for sub-016) but AFNI also flagged sub-016 as having the task condition and not the control condition mildly correlated to motion. These maps provide evidence that task-correlated motion affected data quality.

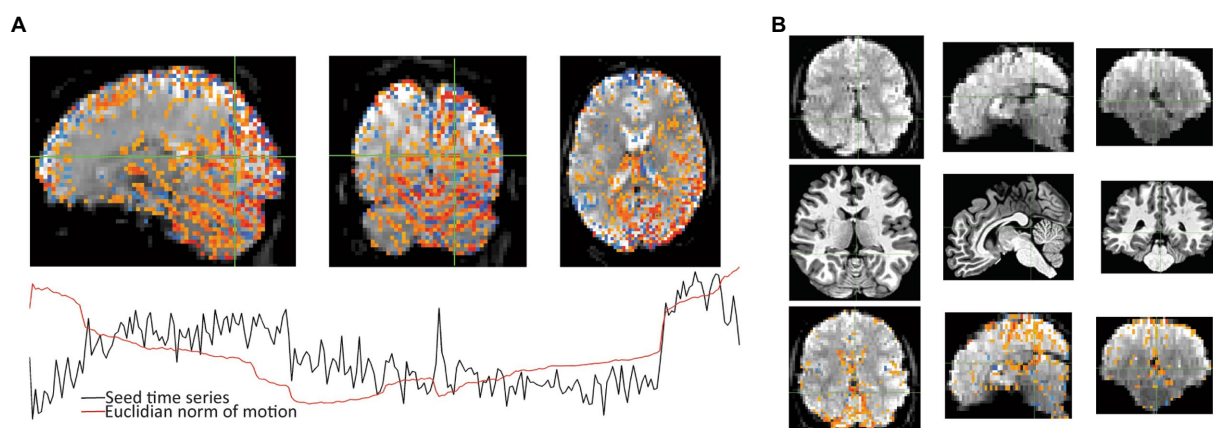


FIGURE 5

After seeing warnings due to “extent of local correlation” and “EPI variance lines” in AFNI’s automatic QC, *instacorr* was used to examine more closely. (A) For the correlation seed at the crosshair, Sub-018, shows an artifactual pattern of correlations ($p < 0.001$) across large portions of the posterior cortex and cerebellum. Time series shows that some of this follows several large jumps in motion. (B) For Sub-002, an unusually large hypointensity was noticed in the unprocessed EPI that was alarming during the initial review. Anatomical viewing of the same slices shows a slightly large superior cistern and 4th ventricle. Correlations to the cross hairs on the unpressed image ($p < 0.001$ with translucency below threshold) shows slightly larger correlations to CSF in the interhemispheric fissure. This observation will likely not cause problems for univariate statistical tests, but it could cause analysis issues if ROIs include this larger area of CSF that contains some internal correlations.

intended purpose until alignment was fixed. If we planned to use these data for a larger study, we would have tried to fix the alignment, but for this demonstration, we ended by noting the alignment issues. This

occurrence highlights how human interpretation is a fundamental part of QC and understanding why data are low quality is sometimes more important than merely identifying low quality data. Even while

TABLE 1 QC Classifications for all subjects.

Subject ID	Status	Notes
sub-001	Include	Dropout in temporal lobe
sub-002	Include	Dropout in temporal lobe
sub-003	Unsure	Dropout in temporal lobe; larger corr between white matter ROI & gray matter; 10.3% vols censored; instacorr showing potentially serious motion artifacts
sub-004	Include	Dropout in temporal lobe
sub-005	Unsure	Dropout in temporal lobe; larger corr between WM & GM; Instacorr shows several widespread artifacts, possibly respiratory
sub-006	Include	Dropout in temporal lobe; larger corr between WM & GM
sub-007	Include	Dropout in temporal lobe
sub-008	Include	Dropout in temporal lobe
sub-009	Exclude	35% vols censored; very large corr between WM & GM; activation hotspots outside of brain
sub-010	Include	Dropout in temporal lobe; larger corr between WM & GM
sub-011	Include	Dropout in temporal lobe
sub-012	Exclude	<i>Instacorr</i> showing some artifacts; 12.8% vols censored; Full F stat map hotspots outside of brain and speckled inside brain; Dropout in temporal lobe
sub-013	Exclude	10.3% vols censored, task vols more censored than control; Full F stat map hotspots outside of brain and speckled inside brain; <i>instacorr</i> showing some artifacts; Dropout in temporal lobe
sub-014	Include	Dropout in temporal lobe; larger corr between WM & GM
sub-015	Include	Larger corr between WM & GM
sub-016	Exclude	Dropout and distortion in temporal and frontal lobes affecting alignment; activation hotspots in CSF; task correlation to motion
sub-017	Exclude	40% vols censored; Dropout in temporal lobe
sub-018	Exclude	<i>Instacorr</i> showed nontrivial MRI artifact correlations
sub-019	Include	Dropout in temporal lobe; larger corr between WM & GM
sub-020	Include	Dropout in temporal lobe
sub-021	Include	12.4% vols censored; Dropout in temporal lobe and cerebellum
sub-022	Exclude	<i>Instacorr</i> showed nontrivial MRI artifact correlations; 17.8% vols censored; Full F stat map speckled inside brain; Dropout in temporal lobe
sub-023	Exclude	Hotspots of activity outside of brain and little robust in-brain hotspots; very large corr between WM & GM; radial corr map shows probably artifacts; 14.9% vols censored; Dropout in temporal lobe
sub-024	Exclude	33.5% vols censored
sub-025	Exclude	15.3% vols censored; task-correlated motion; more motion censoring in task vs. control; very large corr between WM & GM
sub-026	Unsure	19.4% vols censored; larger corr between WM & GM; Dropout in temporal lobe; slightly more censored vols in task vs. control
sub-027	Exclude	19.4% vols censored; Hotspots of activity outside of brain and little robust in-brain hotspots; very larger corr between WM & GM; Dropout in temporal lobe
sub-028	Include	7.9% vols censored
sub-029	Exclude	20.2% vols censored
sub-030	Unsure	<i>Instacorr</i> and local corr maps showed localized artifacts that might require exclusion depending on areas of research interest
sub-101	Exclude	Likely Left/right flip; 20.5% vols censored
sub-102	Include	5.8% vols censored
sub-103	Include	2.6% vols censored; <i>instacorr</i> correlations not great, but nothing clearly exclusionary
sub-104	Include	16% vols censored; <i>instacorr</i> correlations not great, but nothing clearly exclusionary
sub-105	Include	11.5% vols censored; <i>instacorr</i> correlations not great, but nothing clearly exclusionary
sub-106	Exclude	13.5% vols censored; Very large global correlations to seeds
sub-107	Include	19.2% vols censored
sub-108	Include	4.5% vols censored
sub-109	Include	3.8% vols censored

(Continued)

TABLE 1 (Continued)

Subject ID	Status	Notes
sub-110	Include	4.5% vols censored
sub-111	Exclude	7.7% vols censored; Very large global correlations to seeds
sub-112	Include	6.4% vols censored
sub-113	Include	0.6% vols censored
sub-114	Exclude	4.7% vols censored; Very large global correlations or anti-correlations to seeds
sub-115	Exclude	Likely left/right flip
sub-116	Exclude	Neither left–left nor left/right flip is great. With close inspection, unclear if anatomical is same brain as EPI
sub-117	Include	1.9% vols censored
sub-118	Exclude	30.1% vols censored
sub-119	Include	10.9% vols censored
sub-120	Include	1.3% vols censored; <i>instacorr</i> correlations not great, but nothing clearly exclusionary

Task subjects are 001–030, rest subjects are 101–120. Notes explain why a subject was excluded or unsure or highlight something worth continued monitoring in included subjects.

emphasizing the importance of human interpretation, we leaned too heavily on an automated summary image to reject an alignment. This is a critical point since, as study sample sizes increase and data rejection is automated and not followed up by human interpretation, the more likely usable data will be automatically rejected and systematic issues underlying data rejection will be overlooked.

Automated measures combined with human interaction and judgement were essential to the QC process. While automated measures such as correlations to a white matter ROI, statistical result maps, and line variance warnings mandated closer attention, it was direct inspection of volumes and time series, including with using *instacorr*, that became essential for identifying wide-spread issues that warranted data exclusion. Our initial error with mismatching anatomicals and EPIs also highlights the importance and limits of automated QC for registration. The alignment measures showed bad alignments, but not why. For several subjects, the mismatched volumes were subtle even with a close inspection. AFNI's warning for left–right flips is an example where automation can highlight a serious alignment problem that is also subtle. More innovation in automated metrics to assess alignment quality, such as the demonstrated left/right flipping test, is needed. For example, a *post hoc* analysis of our mismatched processing showed that while the cost functions used for alignment are sensitive to the precise contrasts of the EPI and anatomical volumes, since the anatomicals and EPI images had similar contrasts across the dataset, the cost function values for the mismatched fits were clearly higher than the good fits in comparison to other subjects in each dataset (Figure 7E). This is a potential new automated metric that could flag concerning alignments for follow-up by human inspection.

At many points in this project, it was clear that hands-on training was essential. The two authors who conducted most of the visual inspect of results have been working with fMRI data for slightly more than a year. Though the more experienced authors gave consistent instructions, it was impossible to give them comprehensive written instructions that covered the range of issues they observed solely within these datasets. For example, there were several cases where anomalies in images, like a line of CSF that was unusually visible in a single slice caused serious concerns during the initial review, but

expert feedback showed it was not a serious problem (Figure 5B). Improving the training of novice neuroimagers was an interactive and iterative process, where they presented concerning observations and the more experienced neuroimagers helped them understand what issues were or were not actual concerns. Over time, they were able to more independently make appropriate QC judgements. Therefore, such training needs to go beyond a lecture and involve mentored examination of actual datasets.

We have endeavored to provide some points of discussion when devising ways to train people in QC and provide a stable framework for creating a process tailored to individual researchers' needs. Teaching best practices for quality control is far beyond the scope of a single manuscript. Since we focus on QC, rather than what to do after QC, we do not substantively discuss MRI artifacts nor ways to reduce certain artifacts through changes in acquisition or analysis. There are existing reviews on fMRI noise and noise reduction (Liu, 2016; Caballero-Gaudes and Reynolds, 2017), but we are not aware of any published reviews or even book chapters that specifically focus on MRI artifacts for fMRI. While recorded lectures and blog posts cover MRI artifacts, learning to understand and interpret fMRI artifacts remains heavily dependent on hands-on training.

Automation remains essential to QC. Appropriate use of automation can be a very important part of both analysis and quality control when paired with human interpretation and rigorous inspection. When steps are properly automated, human induced errors can be reduced, resulting in more consistent and reproducible results across subjects or analyses. Automated pipelines are also more likely to be neatly organized and understandable, with notes integrated into the scripts that run them rather than scattered across an entire project. This can drastically ease the burden of finding important data or tables to inspect. For the QC metrics demonstrated here, head motion, temporal outlier detection, DOF counts and accompanying censoring and warnings were automatic and appeared robust. Flagging of left–right flipping, while only partially automated, proved invaluable as it is a very difficult problem to spot. Additionally, having a report which organizes much of the relevant information in one place to systematically review, saved many personnel hours during the data

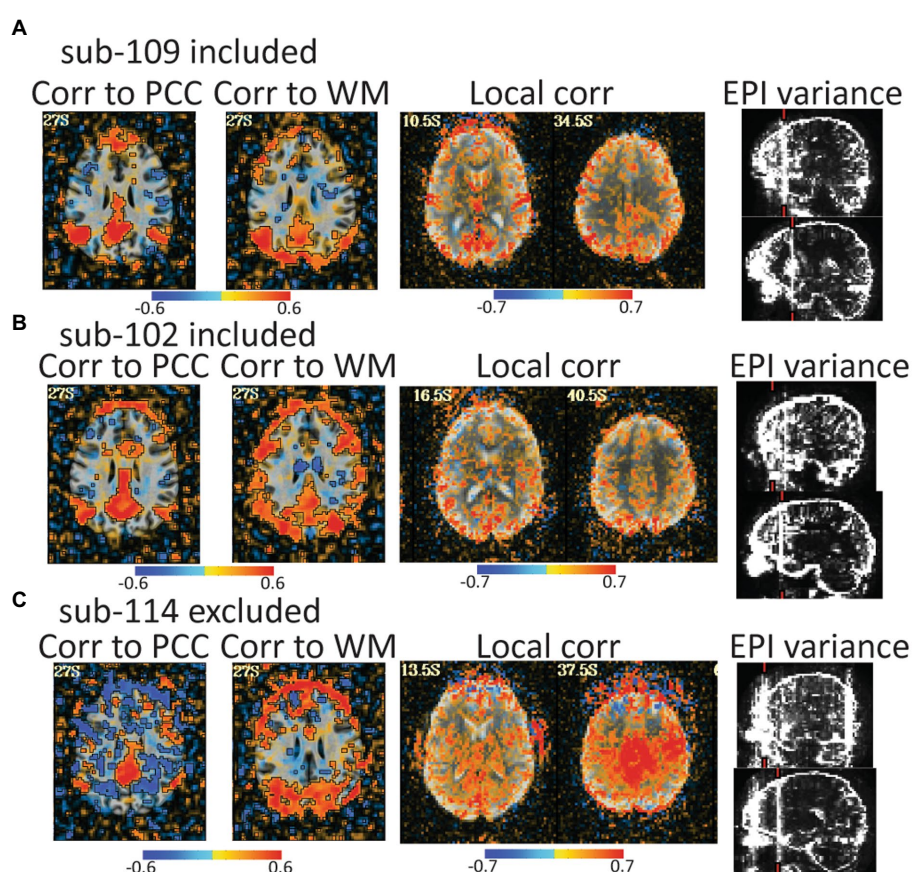


FIGURE 6

Automated QC image from 3 rest data study subjects with low head motion (only 4%–6% of volumes censored). An atlas-based posterior cingulate (PCC) ROI is calculated and the correlation maps (r values), should highlight some default mode network (DMN) connections. Too much correlation between a white matter (WM) ROI and gray matter can be concerning. Local correlations are the correlations of each voxel to surrounding voxels in a 2cm sphere and can highlight scanner artifacts. EPI variance line warnings highlight lines of high variance that might be artifacts. (A) sub-109 has a plausible DMN from the PCC seed, no excessive correlations to white matter, no non-anatomical local correlations, and the variance warnings were checked with *instacorr* and did not show pervasive issues after preprocessing. (B) sub-102 was typical for these data. The DMN is present, but not as clean, there are more WM correlations in and out of the brain, and EPI variance warnings showed some issues with *instacorr*, but not enough to reject. If typical subjects in this dataset were cleaner, we might have rejected sub-102. (C) sub-114 is a clear rejection with non-anatomical anticorrelations to the PCC, large artifacts in WM correlations, a large local correlation, and EPI variance warnings paired with concerning artifacts visible with *instacorr*.

review process and made it possible for the human review to efficiently focus on actual issues.

We used and benefitted from many automated QC measures that are now built-in defaults when AFNI's *afni_proc.py* command is run. Automation is a work in progress and each tool has strengths and weaknesses. We note some places where AFNI's automation can improve under the assumption that these may benefit other tools as well. In particular, connections between reports and the underlying data that generated them could be improved so that it could be easy to quickly navigate to from a concerning image, such as an image of a few slices with questionable alignment, to explore the full alignment in more depth. Another gap in AFNI's automated measures is that there are few automated summaries of QC measures across participants.

The publication describing *MRIQC* tools discusses potential inconsistencies by basing too many decisions on human judgements and recommends a push toward more automated measures (Esteban et al., 2017). While we agree automated measures are essential and they acknowledge human judgement is still important, we think there

can be dangers from over automation or excessive trust in automated thresholds for QC metrics. Automated measures can suffer biases of omission. For example, the lack of automated measures for alignment quality is paired with the lack of a field-wide discussion on the noise and reproducibility issues due to sub-optimal alignments. Automated measures that reject data without human interpretation can also mask underlying and solvable issues.

We believe it is imperative to continue discussing QC priorities, processes, standards, and tools. Moreover, discussions of reproducibility and reliability of fMRI data need to go beyond concerns over head motion and precise yet arbitrary statistical thresholds. Focusing just on one QC concern, like head motion, is like a building inspector looking for signs of water damage. Water damage can be a serious issue and expertise is required to know how to look for such damage, but there is a risk to over-focusing on water damage and missing signs that the floor is about to collapse. Good quality control requires a more comprehensive assessment. The neuroimaging community can do more to understand the full range of problems that exist in data today, so that we can get better at identifying and

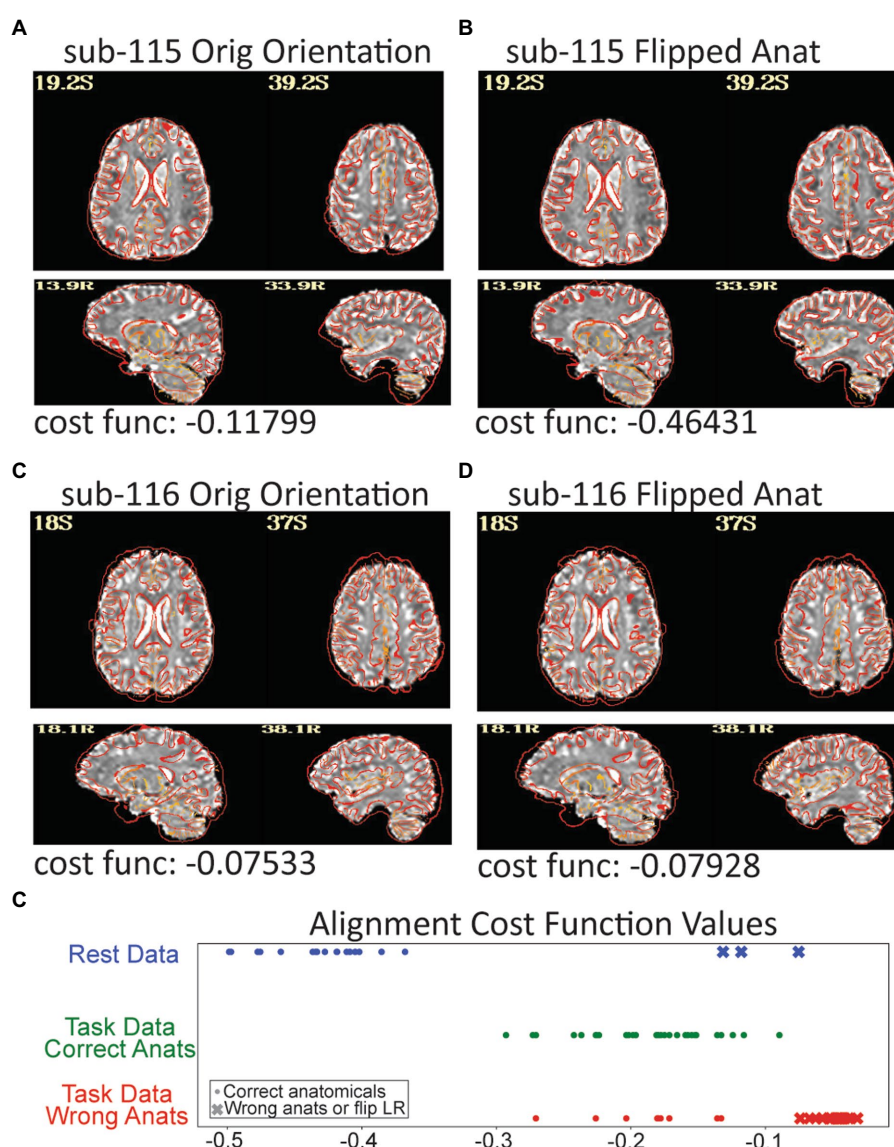


FIGURE 7

Three subjects in the resting data triggered a left–right flip warning which happens when the cost function for anatomical to EPI alignment finds a better local minimum after flipping the anatomical. The grayscale EPI image used for alignment is shown with the edges of the aligned anatomicals. (A) The original alignment for sub-115 looks ok, but (B) shows the alignment for sub-115 with the anatomical image flipped and the gyral edges are clearly better matched. Sub-115 generated a “severe” left–right flip warning. Sub-116 does not have a great alignment for the original (C) or flipped (D) anatomical and generated a “medium” left–right flip warning. Since neither fits well, sub-116 may have been shared with the wrong anatomical image. (E) The cost function minimums for the successful alignments in the rest dataset were -0.36 to -0.5 while the 3 flipped alignments were more than -0.13 . Similarly, when the task data were unintentionally aligned to the wrong anatomicals, the cost functions were much higher. While cost functions are relative measures, the values may be useable as an intra-study alignment QC measure.

documenting problems. Particularly as data sharing becomes the norm, the more we can do to improve QC processes today, the more likely our current data will still be useful for future research.

We have demonstrated a typical QC process for our research group. We have likely missed some data quality problems that other researchers may catch because processes vary and are often tailored to different research approaches. This is one of the reasons we highlight the importance of the underlying scientific questions and context for good QC. We hope more researchers will share their QC protocols, so that a wide array of approaches can be compared and used to improve the next generation of QC tools and processes.

6. Conclusion

Good data quality is essential for reproducible science. Quality control processes help validate data quality and ensure data are suitable to address experimental questions. Timely QC steps during the early stages of a study can improve data quality and save resources by identifying changes to acquisitions or analyses that can address problems that arise during QC. QC is an ongoing process that does not end after the early stages of a study. Shared data are not inherently quality-checked data, and even shared data that includes a documented QC process and output may not

be sufficient since priorities for quality checks can be study context-dependent.

A good QC process should be integrated into study planning. While automation should be used wherever possible, human observations and interpretations are critical. Much discussion of QC focuses on the binarized decision of whether to keep or exclude data, but we find that a key element of QC is to identify potentially correctable issues. Particularly, as fMRI studies increase in size or aggregate multiple datasets, good QC processes will require planning that includes decisions on what can be automated and what will require peoples' time.

Much public discussion about reproducible neuroimaging has focused on appropriate sample sizes, statistical tools, and thresholds. We posit that normalizing timely and rigorous QC is an equal if not more important step our field can take to improve reproducibility. While we present a framework for thinking about fMRI QC along with a demonstration of one existing QC pipeline on a couple of shared datasets, this is far from sufficient. Quality control priorities and methods deserve more attention, discussion, and innovation from the neuroimaging community.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: <https://osf.io/qaesm/>.

Ethics statement

The studies involving human participants were reviewed as part of this special issue. The editors selected openly available ethics committee approved datasets, which we analyzed without knowing the source of the data. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JT, DH, JG-C, and PB contributed to overall study design. JT wrote the processing scripts and guidelines for running and checking the exemplar data. MH, MS, and DH processed most data and did the quality checks under the direct supervision of JT with active support and guidance from DH and JG-C. DH and JT wrote the manuscript with some figures by MS and MH. All authors gave feedback on the

manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the Intramural Research Program of the NIMH under grant number ZIAMH002783.

Acknowledgments

We would like to thank Peter J. Molfese, Tyler Morgan, and Sharif Kronemer for their helpful comments. We thank Paul Taylor for actively prompting us to figure out why so many of our initial alignments failed. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). The scientists who publicly shared the data used in these demonstrations made this work possible. The views expressed in this article do not necessarily represent the views of the National Institutes of Health, the Department of Health and Human Services, or the United States Government.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2023.1100544/full#supplementary-material>

References

- Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. *NeuroImage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034
- Birn, R. M., Murphy, K., Handwerker, D. A., and Bandettini, P. A. (2009). fMRI in the presence of task-correlated breathing variations. *NeuroImage* 47, 1092–1104. doi: 10.1016/j.neuroimage.2009.05.030
- Caballero-Gaudes, C., and Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage* 154, 128–149. doi: 10.1016/j.neuroimage.2016.12.018
- Caparelli, E. C., Ross, T. J., Gu, H., and Yang, Y. (2019). Factors affecting detection Power of blood oxygen-level dependent signal in resting-state functional magnetic resonance imaging using high-resolution Echo-planar imaging. *Brain Connect.* 9, 638–648. doi: 10.1089/brain.2019.0683
- Cheng, C., and Halchenko, Y. (2020). A new virtue of phantom MRI data: explaining variance in human participant data [version 1; peer review: awaiting peer review]. *F1000research* 9:1131. doi: 10.12688/f1000research.24544.1
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Cox, R. W., and Jesmanowicz, A. (1999). Real-time 3D image registration for functional MRI. *Magn. Reson. Med.* 42, 1014–1018. doi: 10.1002/(SICI)1522-2594(199912)42:6<1014::AID-MRM4>3.0.CO;2-F
- Dosenbach, N. U. F., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., et al. (2017). Real-time motion analytics during brain MRI improve data

- quality and reduce costs. *NeuroImage* 161, 80–93. doi: 10.1016/j.neuroimage.2017.08.025
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* 12:e0184661. doi: 10.1371/journal.pone.0184661
- Fischl, B., and Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11050–11055. doi: 10.1073/pnas.200033797
- Friedman, L., and Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *J. Magn. Reson. Imaging* 23, 827–839. doi: 10.1002/jmri.20583
- Friedman, L., Glover, G. H., and Fbirn, C. (2006). Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *NeuroImage* 33, 471–481. doi: 10.1016/j.neuroimage.2006.07.012
- Glen, D. R., Taylor, P. A., Buchsbaum, B. R., Cox, R. W., and Reynolds, R. C. (2020). Beware (surprisingly common) left-right flips in your MRI data: an efficient and robust method to check MRI dataset consistency using AFNI. *Front. Neuroinform.* 14:18. doi: 10.3389/fninf.2020.00018
- Gotts, S. J., Saad, Z. S., Jo, H. J., Wallace, G. L., Cox, R. W., and Martin, A. (2013). The perils of global signal regression for group comparisons: a case study of autism Spectrum disorders. *Front. Hum. Neurosci.* 7:356. doi: 10.3389/fnhum.2013.00356
- Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., et al. (2018). Behavioral interventions for reducing head motion during MRI scans in children. *NeuroImage* 171, 234–245. doi: 10.1016/j.neuroimage.2018.01.023
- Heunis, S., Lamerichs, R., Zinger, S., Caballero-Gaudes, C., Jansen, J. F. A., Aldenkamp, B., et al. (2020). Quality and denoising in real-time functional magnetic resonance imaging neurofeedback: a methods review. *Hum. Brain Mapp.* 41, 3439–3467. doi: 10.1002/hbm.25010
- Huber, L. R., Poser, B. A., Bandettini, P. A., Arora, K., Wagstyl, K., Cho, S., et al. (2021). LayNii: a software suite for layer-fMRI. *NeuroImage* 237:118091. doi: 10.1016/j.neuroimage.2021.118091
- Huguet, J., Falcon, C., Fusté, D., Girona, S., Vicente, D., Molinuevo, J. L., et al. (2021). Management and quality control of large neuroimaging datasets: developments from the Barcelonaβeta brain research center. *Front. Neurosci.* 15:633438. doi: 10.3389/fnins.2021.633438
- Jo, H. J., Saad, Z. S., Simmons, W. K., Milbury, L. A., and Cox, R. W. (2010). Mapping sources of correlation in resting state FMRI, with artifact detection and removal. *NeuroImage* 52, 571–582. doi: 10.1016/j.neuroimage.2010.04.246
- Kim, H., Irimia, A., Hobel, S. M., Poghosyan, M., Tang, H., Petrosyan, P., et al. (2019). The LONI QC system: a semi-automated, web-based and freely-available environment for the comprehensive quality control of neuroimaging data. *Front. Neuroinform.* 13:60. doi: 10.3389/fninf.2019.00060
- Liu, T. T. (2016). Noise contributions to the fMRI signal: an overview. *NeuroImage* 143, 141–151. doi: 10.1016/j.neuroimage.2016.09.008
- Liu, T. T., Glover, G. H., Mueller, B. A., Greve, D. N., Rasmussen, J., Voyvodic, J. T., et al. (2015). “Quality Assurance in Functional MRI” in *fMRI: From nuclear spins to brain functions biological magnetic resonance*. eds. K. Uludag, K. Ugurbil and L. Berliner (Boston, MA: Springer), 245–270.
- Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., et al. (2013). Human connectome project informatics: quality control, database services, and data visualization. *NeuroImage* 80, 202–219. doi: 10.1016/j.neuroimage.2013.05.077
- Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., et al. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* 20, 299–303. doi: 10.1038/nn.4500
- Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., and Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage* 40, 409–414. doi: 10.1016/j.neuroimage.2007.11.048
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Scott, C. J. M., Arnott, S. R., Chemparathy, A., Dong, F., Solovey, I., Gee, T., et al. (2020). An overview of the quality assurance and quality control of magnetic resonance imaging data for the Ontario neurodegenerative disease research initiative (ONDRI): pipeline development and neuroinformatics. *Biorxiv*, 2020.01.10.896415 [Preprint]. doi: 10.1101/2020.01.10.896415
- Strand, J. F. (2023). Error tight: exercises for lab groups to prevent research mistakes. *Psychol. Methods*. doi: 10.1037/met0000547
- Taylor, C. M., Pritschet, L., Olsen, R. K., Layher, E., Santander, T., Grafton, S. T., et al. (2020). Progesterone shapes medial temporal lobe volume across the human menstrual cycle. *NeuroImage* 220:117125. doi: 10.1016/j.neuroimage.2020.117125
- Wang, R. Y., and Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *J. Manag. Inform. Syst.* 12, 5–33. doi: 10.1080/07421222.1996.11518099
- Yang, G. J., Murray, J. D., Repovs, G., Cole, M. W., Savic, A., Glasser, M. F., et al. (2014). Altered global brain signal in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* 111, 7438–7443. doi: 10.1073/pnas.1405289111



OPEN ACCESS

EDITED BY

Richard Craig Reynolds,
Clinical Center (NIH), United States

REVIEWED BY

Jennifer Evans,
National Institutes of Health (NIH),
United States
Martha J. Holmes,
University of Cape Town, South Africa

*CORRESPONDENCE

Rebecca J. Lepping
✉ rlepping@kumc.edu

RECEIVED 22 October 2022

ACCEPTED 07 April 2023

PUBLISHED 04 May 2023

CITATION

Lepping RJ, Yeh H-W, McPherson BC,
Brucks MG, Sabati M, Karcher RT, Brooks WM,
Habiger JD, Papa VB and Martin LE (2023)
Quality control in resting-state fMRI: the
benefits of visual inspection.
Front. Neurosci. 17:1076824.
doi: 10.3389/fnins.2023.1076824

COPYRIGHT

© 2023 Lepping, Yeh, McPherson, Brucks,
Sabati, Karcher, Brooks, Habiger, Papa and
Martin. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Quality control in resting-state fMRI: the benefits of visual inspection

Rebecca J. Lepping^{1,2*}, Hung-Wen Yeh^{3,4}, Brent C. McPherson⁵,
Morgan G. Brucks^{2,6}, Mohammad Sabati^{2,7}, Rainer T. Karcher²,
William M. Brooks^{1,2}, Joshua D. Habiger⁸, Vlad B. Papa² and
Laura E. Martin^{2,6}

¹Department of Neurology, University of Kansas Medical Center, Kansas City, KS, United States,

²Hoglund Biomedical Imaging Center, University of Kansas Medical Center, Kansas City, KS, United States, ³Division of Health Services and Outcomes Research, Department of Pediatrics, Children's Mercy Research Institute, Kansas City, MO, United States, ⁴Department of Pediatrics, School of Medicine, University of Missouri-Kansas City, Kansas City, MO, United States, ⁵Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada, ⁶Department of Population Health, University of Kansas Medical Center, Kansas City, KS, United States, ⁷Bioengineering Program, School of Engineering, University of Kansas, Lawrence, KS, United States, ⁸Department of Statistics, Oklahoma State University, Stillwater, OK, United States

Background: A variety of quality control (QC) approaches are employed in resting-state functional magnetic resonance imaging (rs-fMRI) to determine data quality and ultimately inclusion or exclusion of a fMRI data set in group analysis. Reliability of rs-fMRI data can be improved by censoring or “scrubbing” volumes affected by motion. While censoring preserves the integrity of participant-level data, including excessively censored data sets in group analyses may add noise. Quantitative motion-related metrics are frequently reported in the literature; however, qualitative visual inspection can sometimes catch errors or other issues that may be missed by quantitative metrics alone. In this paper, we describe our methods for performing QC of rs-fMRI data using software-generated quantitative and qualitative output and trained visual inspection.

Results: The data provided for this QC paper had relatively low motion-censoring, thus quantitative QC resulted in no exclusions. Qualitative checks of the data resulted in limited exclusions due to potential incidental findings and failed pre-processing scripts.

Conclusion: Visual inspection in addition to the review of quantitative QC metrics is an important component to ensure high quality and accuracy in rs-fMRI data analysis.

KEYWORDS

artifacts, functional magnetic resonance imaging (fMRI), resting state—fMRI, reproducibility of results, quality control

Introduction

Quality control (QC) in functional magnetic resonance imaging (fMRI) data is a critical step in ensuring accurate interpretation of results and reliable and replicable findings. Data may be corrupted at acquisition due to hardware or software malfunctions, artifacts from metallic objects, spurious physiological signals (heart rate, respiration, etc.) or participant motion. Further, incidental findings of atypical anatomic formations, lesions, or other injury may

be grounds for data exclusion if those findings are related to inclusion and exclusion criteria for the study, or if they cause errors in certain processing steps. There is a clear need for consensus on QC approaches for fMRI data, and for a revisiting of reporting standards to improve cross-study interpretation and replicability (Esteban et al., 2017). There are emerging approaches to crowd-source the QC of imaging data sets using a combination of expert curation and a gamified interface for identifying scans. In this paper, we describe our fMRI QC methods from data acquisition through individual preprocessing. Our methods rely on standard tools available through the analysis software we use and also include visual inspection by trained reviewers at multiple stages of the process. While the field recognizes the value of quantitative metrics and automated processes for evaluating data quality, we believe there is added value in qualitative assessment that cannot be captured by quantitative measures of displacement, censoring, or signal intensity or homogeneity. We apply these QC strategies to a publicly available data set and report out standardized outcomes identified in the Frontiers Research Topic, Demonstrating Quality Control (QC) Procedures in fMRI.

Across MRI imaging protocols, fMRI data are particularly sensitive to participant head motion. Strategies exist to minimize participant head motion at data acquisition, such as the use of foam padding around the head, a strap across the forehead, bite bars, or real-time feedback to the participant and prospective motion corrections (Thulborn, 1999; Lazar, 2008; Vanderwal et al., 2015). However, these often require specialized settings, sequences, or equipment and are not sufficient to eliminate all movement and some data will be lost to motion corruption.

One of the most observable effects of head motion on fMRI data is the increase or decrease in signal in the affected volumes. In the case of blood oxygen level dependent (BOLD) imaging, data are acquired in slices through the volume of the brain over the course of a few seconds. The slice to be imaged is excited with a radio-frequency (RF) pulse, and the echo is read out a few milliseconds later. If the excited slice has moved in space, the echo will not be accurately read, leading to reduced signal in that slice. Additionally, the next slice to be acquired may have been excited by the preceding pulse and may have residual signal. A second RF pulse in that slice would lead to increases in signal readout. For these reasons, the volumes surrounding a motion spike are often also unreliable, and these effects may last for several seconds (Power et al., 2014). Compounding this issue is that all voxels within a slice or volume are not likely to be impacted the same way, as motion is rarely limited to translation along a single axis. Because of this, the relationship between signal within a given voxel and motion parameters is not linear (Power et al., 2015). Motion can decrease the fMRI signal temporal stability by causing signal alterations across volumes which eventually increase false outcomes (Satterthwaite et al., 2013). Moreover, motion can potentially modulate connectivity-related measurements because it produces global signal changes resulting in spurious results (Rogers et al., 2007).

Certain populations may be especially prone to movement during fMRI scanning. Children, older people, people with back pain, or people with high impulsivity may not be able to hold still for an entire functional scan, which can last for several minutes (Fox and Greicius, 2010; Couvy-Duchesne et al., 2014; Kong et al., 2014; Couvy-Duchesne et al., 2016; Pardoe et al., 2016). Therefore, the development of new approaches and the optimization of current strategies to reduce

motion-related artifacts in fMRI data sets are critical for imaging studies of these populations. Because resting state correlation relies on low frequency modulation within the signal, longer scans are recommended (up to ~10 min in some cases) (Birn et al., 2013), potentially exacerbating the problem of participant movement. Participants may tolerate several shorter scans with breaks in between – collecting multiple resting state scans and concatenating across them improves the signal-to-noise ratio (Chen et al., 2010); however, no strategy completely eliminates the impacts of participant head motion (Power et al., 2014, 2015).

The statistical approach of including motion parameters as nuisance regressors in the analysis reduces the impact of motion and has been widely adopted as a standard processing step (Johnstone et al., 2006). It has been shown that removing, or censoring, only the volumes most affected by motion prior to statistical analysis improves reliability (Power et al., 2012; Carp, 2013; Power et al., 2013). Additional ‘scrubbing’ or removing physiological noise signals is also helpful for removing spurious correlations due to head motion (Siegel et al., 2014). However, censoring alone still has problems. One is how to choose the optimal censoring threshold, which may depend on the level of motion in a data set (Power et al., 2014). Once a threshold has been chosen, another concern is that correlation estimates from participants with reduced data sets after censoring may be noisy or have extreme values that may influence group statistics or reduce power. To address this, many studies also exclude entire participants or scans that exceed pre-specified censoring limits (Power et al., 2015). Excluding participants with greater than 10% censored is often used as a threshold, and less conservative censoring thresholds of 15–25% have been used with pediatric populations (Siegel et al., 2014). An entirely different approach from censoring is to use independent components analysis (ICA) to identify the signal associated with head motion (Griffanti et al., 2014; Siegel et al., 2014; Patriat et al., 2015, 2016; Pruim et al., 2015). Since reliability is dependent on the length of usable data, some researchers exclude participants with usable resting state scan data less than ~5 min after censoring (Van Dijk et al., 2012; Andellini et al., 2015).

Censoring or scrubbing solutions allow for removing motion corrupted data while preserving some data and avoiding excluding entire participant data sets. If motion corruption causes data to not be missing at random, excluding more data in one group than another can cause bias in estimation and result in loss in power or invalid testing procedures (Little and Rubin, 2002). Moreover, excluding acquired data introduces a waste of resources and excessive costs for research services and personnel time. Given the challenge of recruiting well-characterized participants from clinical populations, the commitment of participants, and the cost of data collection and analytic staff, there are financial and social burdens to unnecessarily excluding data.

While motion artifacts have been well-documented to lead to both type I and type II errors in downstream analyses, other issues can and do arise during functional data acquisition and analysis. These include incidental findings of anatomic variability in the images which could indicate a medical concern or a benign anatomic difference that is of little medical concern. These findings, however, could be reason for participant exclusion, for example there is an incidental finding that indicates a previous stroke and stroke is an exclusion criterion for the study. Also, these anatomic variabilities could lead to issues with misalignment or normalization into template space, therefore, visual

inspection of the results is warranted. In addition to anatomic variabilities and incidental findings, script failures are another source of error in rs-fMRI data analysis. Analysis of rs-fMRI data is performed as a series of steps, with each step taking the output from the previous step, performing another process, and then generating a new output image. Errors are possible at each step, and it is critical to determine that scripts are performing correctly so that the input–output–input chain does not result in errors in the final output data set. These errors are sometimes difficult to find if one only examines quantitative QC metrics, but can be easy to assess visually, for example if the entire functional series of images have been flipped upside down during processing but are centered with the anatomic image, global metrics of homogeneity will not differ between a correctly aligned and incorrectly aligned image. If such processing errors are allowed into group analysis, the spatial location of anatomy will not match across all participants.

In this paper, we describe our processes for rs-fMRI QC, including review of quantitative and qualitative software-generated metrics and visual inspection at each processing step to ensure the most accurate data are carried forward in the analysis process. Further, we advocate for including as much data as possible to minimize bias and honor the participant time and research resources provided.

Materials and methods

We performed an analysis of previously published and publicly available human participants' data provided as part of the Demonstrating QC Procedures in fMRI Research Topic (Biswal et al., 2010; Di Martino et al., 2014; Markiewicz et al., 2021). Briefly, resting state fMRI (rs-fMRI) data were pulled from publicly available datasets (ABIDE, ABIDE-II, functional Connectome Project, Open Neuro) across seven imaging sites, with approximately 20 participants from each site. Imaging parameters are summarized in Table 1. For this Research Topic, the Project leaders renamed the data with new participant IDs and organized them into BIDS common directory format. Each participant had one anatomical image and one or two rs-fMRI sequences. Imaging parameters for the rs-fMRI sequences are reported in Table 1. No information on participant demographics or other characteristics was provided. For the remainder of the paper, we will refer to this data set as the “QC data set.” All procedures involving human participants were performed in accordance with the ethical standards of the Declaration of Helsinki, and the study was approved by the Institutional Review Board where the data were collected. Informed consent was obtained from all participants.

Data processing

MRI data preprocessing and statistical analyses took place in Analysis of Functional Neuroimages (AFNI v22.1.10) (Cox, 1996) and implemented using `afni_proc.py` (Example 11b). Anatomical data were skull stripped and normalized to standard Montreal Neurological Institute (MNI) space using non-linear warping with AFNI command `@SSwarper` and these parameters were applied to the functional data for spatial normalization. The first two volumes of the functional scans were removed, and data were despiked. Volumes were slice time corrected and co-registered to the minimum outlier within the run.

Volumes where more than 5% of the brain voxels were considered outliers and were removed from the analysis. In addition, volumes with motion greater than 0.2 mm within a volume were censored and removed from the analysis. Nuisance variables included motion parameters (3 translation, 3 rotation), average ventricle signal, and average white matter signal. Ventricle signals were estimated by combining an MNI ventricle mask with the participant's cerebral spinal fluid mask derived from the anatomic images. Using multiple regression, a residual time series was calculated for each voxel. The residual time series was then smoothed with a 4 mm FWHM Gaussian kernel, resampled to a $2.5 \times 2.5 \times 2.5$ mm grid, and transformed to MNI space.

Quality control process

Data quality was determined using a combination of quantitative metrics and qualitative assessment (Figure 1). Quantitative metrics included verification of final voxel resolution and outputs from AFNI's APQC of average motion per TR, max motion displacement, and censor fraction. Quantitative metrics were recorded in our REDCap QC checklist (see supplement) for ease of summary and comparison across participants.

Qualitatively, data were viewed by trained staff who made inclusion/exclusion decisions. Training of staff included walking through each step of our REDCap QC checklist (see supplement) and implementing a double data check system where new staff and trained staff both check and verify the same datasets. Staff were considered trained after inclusion/exclusion decisions were consistent with those made by trained/established staff. This method is a step-by-step approach to reviewing data and documenting the results of each of these steps utilizing a standardized REDCap form. This approach is easy to train new raters – we have successfully trained people across all levels of education, from high school students to those with PhDs – and the double-data entry step facilitates inter-rater reliability assessment. Data entry into REDCap also allows summary data to be easily compiled across participants, and if the checklist is used across multiple studies, data can be easily compared across projects. The inclusion of image examples of poor quality data within our REDCap checklist should improve the replicability and inter-rater reliability as well.

Raw DICOM files were converted to NIFTI format prior to being shared publicly, however, when starting from raw DICOM files, our QC process begins with a verification of data completeness comparing file count, file size, and image acquisition parameters against study protocols. We downloaded the NIFTI files and scrolled through the brain slice by slice within AFNI in order to assess each modality for any acquisition issues, distortion of images, or incidental findings. `@SSwarper` outputs were visually inspected for good alignment (clear match between the skull-stripped brain and the MNI base template space) and skull-stripping (little to no clipped/missing brain data) prior to being processed with the individual data set `afni_proc.py` script. We then followed AFNI's standard processing guidelines to check the processed data using the `afni_proc.py` quality control output. REDCap QC included checking the APQC and recording of following: excessive motion, warping, and distortion of the original data, alignment issues between the epi to anatomy and anatomy to the MNI template, inspection of the statistics volumes for excessive noise within

TABLE 1 Resting state fMRI imaging parameters from the seven imaging sites.

Site	Scanner	Field strength	Orientation	In-plane resolution	Spacing between Slices	Repetition time (TR)	Echo time (TE)	Number of Slices	Number of volumes	Parallel reduction (Yes/No)
1	Phillips Achieva	3T	Axial	2.67 mm × 2.67 mm	3.0 mm	2,500 ms	30 ms	47	156	Yes
2	Phillips Achieva	3T	Axial	3.0 mm × 3.0 mm	3.84 mm	2,000 ms	28 ms	38	150	Yes
3	Phillips Achieva DS	3T	Axial	2.56 mm × 2.56 mm	3.1 mm	2,500 ms	30 ms	45	162	Yes
4	Unknown	3T	Unknown	2.67 mm × 2.67 mm	3.0 mm	2,500 ms	Unknown	47	123	Unknown
5	Phillips Achieva OR Siemens TrioTim OR Siemens Prisma_fit	3T	Axial	1.88 mm × 1.88 mm/3.0 mm × 3.0 mm/3.0 mm × 3.0 mm	4.0 mm/4.0 mm/4.0 mm	2,000 ms/2000 ms/2,000 ms	34 ms/30 ms/25 ms	Varied 34–39	144/144/144	Unknown
6	Siemens MAGNETOM Trio	3T	Unknown	4.0 mm × 4.0 mm	4.0 mm	2,500 ms	27 ms	32	varied 130–724	Unknown
7	Siemens Verio	3T	Unknown	3.0 mm × 3.0 mm	3.51 mm	2,500 ms	30 ms	39	198	Unknown

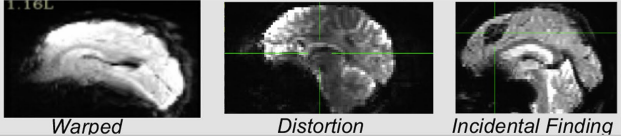
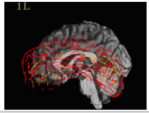
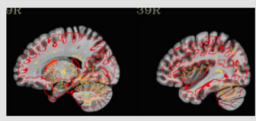

Data Checking Steps		
	Quantitative	Qualitative
1) Check data from scanner	Review file counts	a) Review quality of image (e.g., blurry, warped, distortion) b) Check for potential incidental findings 
2) Check alignment of functional and anatomical scans		Review alignment 
3) Check alignment of anatomical data to template		Review alignment 
4) Check statistics volumes		Look for and note outside of the brain
5) Check motion and outliers	a) Review % censored b) Review max displacement	Review motion plots 
6) Check regressors, degrees of freedom, residuals	Note the degrees of freedom	Note the pattern of activation (e.g., lots, few, unsure)
7) Check warnings	a) Note regression warnings b) Note censor fraction warnings c) Note pre-steady state warnings d) Note left/right flip warnings	
8) Check quantitative data	a) Note final voxel resolution b) Note average motion per TR c) Note number of runs	
9) Decision	a) Summarize information from prior steps b) If data are questionable, review with another member of the study team c) Decide to include or exclude data from final analysis	

FIGURE 1

Data checking steps include qualitative and quantitative evaluation of the imaging data to determine inclusion in group level analysis.

and outside of the brain, excessive motion, low degrees of freedom, warnings, and a brief summary of the @ss_review_basic. Motion and warnings regarding the severity of the overall censor fraction were recorded at three thresholds based on AFNI warning levels (excluding

severe censoring >50%, excluding medium censoring >20%, excluding mild censoring >10%).

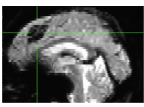
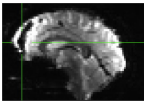
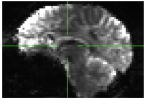
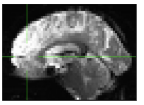
In addition to the steps described above, which follow standard processing guidelines from AFNI, if excessive motion

was present (>20% censoring), we further checked the epi using @epi_review to visually inspect each run. If major alignment or warping issues were present, we used the @ss_review_driver to visualize the data and troubleshoot challenges in the pre-processing steps. This additional visual inspection process may help identify when a script failed and provide visualization of slices that may not be shown in the APQC file. Data were considered usable if there were no incidental findings, if the functional images were clear with little to no warping or blurring, and if the functional images were well aligned with both the anatomic images and the template. Data were excluded if the preprocessing scripts did not successfully complete after three attempts.

Results

Of the 129 available data sets, six data sets were excluded due to (A) Script did not complete successfully ($n=2$), (B) Distortion in the functional image ($n=1$), or (C) Incidental findings ($n=3$; Table 2). No data sets were excluded due to motion, leaving 123 data sets to be included for subsequent analysis (Figure 2). The QC data set contained relatively low levels of motion in terms of quantitative metrics: total censor fraction (Mean=11%, SD=17%) and max displacement (Mean=1.25 mm, SD=0.77 mm). Despite a relatively low censor fraction and max displacement, 30.9% of the data sets had mild censoring or greater (>10%), 14.6% had medium censoring or greater (>20%), and 6.5% of the data sets had severe censoring greater than 50% (Table 3).

TABLE 2 Excluded resting state data sets.

ID	Exclude	QC criteria failed (rationale)	Notes/Examples
315	X	C (incidental finding, black hole in epi file)	
405	X	C (incidental finding, black hole in epi file)	
409	X	B (distortion in the epi file)	
518	X	A (brain was flipped, script failed 3+ times)	During the volume registration step the functional data flipped and problem could not be resolved
519	X	A (brain was flipped, script failed 3+ times)	During the volume registration step the functional data flipped and problem could not be resolved
716	X	C (incidental finding, atrophy and lesions in epi file)	

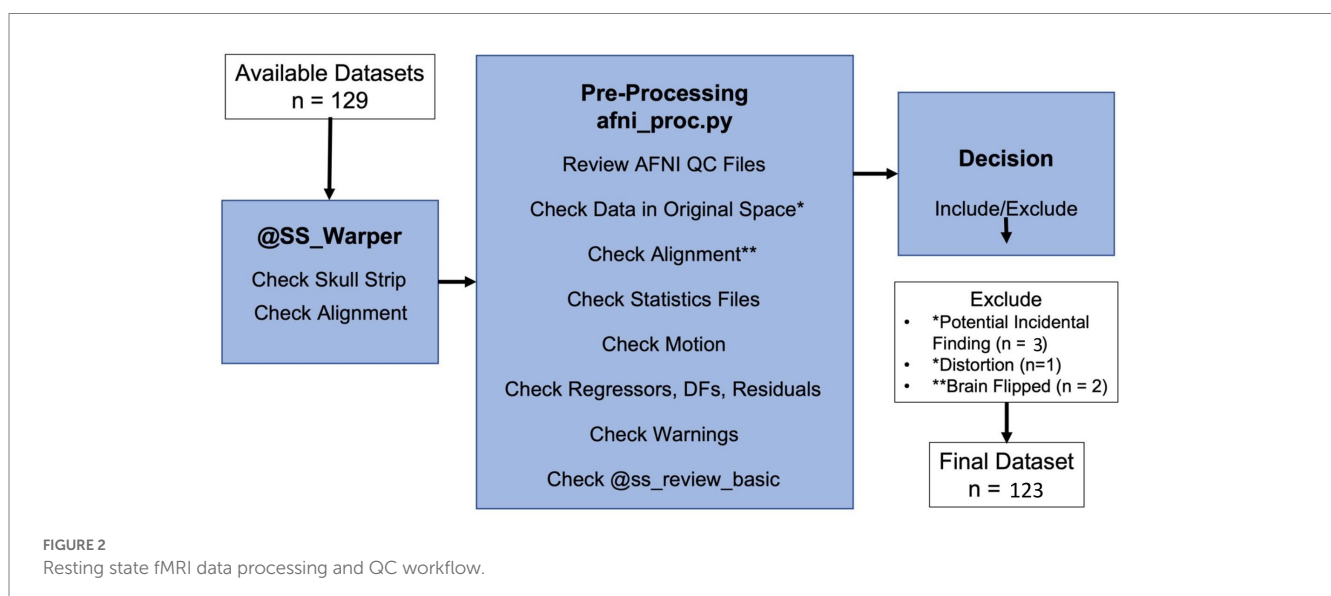


TABLE 3 Resting state data sets exceeding quantitative QC criteria for motion by severity level.

ID	Mild censoring (>10%)	Medium censoring (>20%)	Severe censoring (>50%)
101	X	X	
102	X		
104	X	X	
105	X		
106	X	X	
107	X	X	
109	X		
111	X		
112	X		
114	X		
208	X		
214	X		
307	X	X	X
309	X	X	
314	X	X	
316	X	X	X
402	X		
408	X		
422	X		
502	X		
504	X		
506	X		
507	X	X	X
508	X		
509	X		
511	X	X	
512	X	X	
601	X		
620	X	X	
701	X	X	
703	X	X	X
705	X	X	
706	X	X	X
708	X	X	X
710	X		
712	X	X	X
713	X	X	
714	X	X	X
715	X	X	
Totals	39	21	8

Discussion

The QC approach described above avoids the use of thresholds for excluding participants and favors inclusion of as many data sets as possible and emphasizes qualitative approaches to QC. A variety

of QC approaches can be used to determine data quality and ultimately inclusion or exclusion of a fMRI data set in group analysis, and there are no standards for reporting qualitative approaches. Image artifacts, incidental anatomic findings, and alignment failures that may cause mislocalization of functional data

in anatomic space are our primary reasons for excluding data. These features may be missed if only quantitative metrics are used to evaluate data quality. Global metrics such as homogeneity and censoring are unlikely to vary if an image is flipped upside down or if there is an area of localized hypointensity on the BOLD images indicating a potential incidental finding.

Regarding motion, rather than removing entire data sets from group analysis based on excessive censoring as is commonly done, we advocate for relying on within participant censoring and scrubbing methods to clean motion-related artifacts. There is a non-inclusion aspect as well as real dollar cost when excluding data. Often funded by grants, research money is spent recruiting participants, acquiring data, and paying staff to analyze those data. In addition, participants have volunteered their time into studies. Hence, we as researchers have a social and financial obligation to use the data we have collected to the fullest extent and to get the greatest power out of them that we can. This dataset had relatively little motion; however, nearly 15% would have been excluded had we used a threshold approach at medium (>20%) censoring. We have successfully used this inclusive approach in several studies where motion was a greater concern, including studies in a pediatric population (Lepping et al., 2015, 2019).

Some aspects of motion are more challenging to compensate. Minimizing participant motion at data acquisition is ideal; however, this is not realistic in all situations. Several publications offer methods for prospective motion correction for echo-planar imaging (EPI) (Muraskin et al., 2013; Herbst et al., 2015; Maziero et al., 2020). This is achieved by using an in-scanner camera for head tracking to measure head motion in real time and prospectively adjusting the acquisition positioning accordingly. Other useful methods have been developed for fMRI to adjust acquisition positioning during scanning by measuring and correcting for head motion in real time and prospectively for EPI sequences and with further improvement when combined with retrospective motion correction methods (Lee et al., 1998; Thesen et al., 2000; Beall and Lowe, 2014; Lanka and Deshpande, 2019). While not perfect, some of these prospective methods have been successfully used in resting-state functional connectivity analyses (Lanka and Deshpande, 2019), however, these methods are not available for all researchers, and additional sequence and statistical considerations are still needed.

Many of the imaging analysis software packages have added QC tools that have made it easier to assess data quality and report standard quality metrics across packages. AFNI's APQC html output solidified many of the quality assessment steps we were doing already, including many of the qualitative visual inspection steps. Additional functionality, if provided in the software packages, would further improve the QC process. First, the APQC html file does not currently support saving the data checking within the file itself. Because of this, we have used a separate tool, our REDCap checklist to house the assessments. Second, we have incorporated examples of poor quality data within our REDCap checklist. If that were included in the software output, raters could easily see what the data should not look like, and training for qualitative assessment would be more consistent. Next, other tools within AFNI create QC output files that indicate whether alignment or other downstream steps are likely to fail. Adding that to the APQC process would be useful. Finally, we use the REDCap checklist and project database to export summary QC data for an

entire project. It would be helpful to have a group summary QC output directly from the analysis software.

Conclusion

While quantitative QC metrics including motion are important data to consider when assessing fMRI data quality, some data quality issues may be missed if only quantitative assessments are conducted. Our use of visual inspection throughout the data analysis process ensures that anatomic incidental findings, image artifacts, and processing errors are removed prior to group analysis. Our REDCap checklist can be used to facilitate training of staff and reporting image quality.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: <https://osf.io/qaesm/wiki/home/>. Analysis scripts used in this manuscript can be found here: <https://github.com/rlepping/kumc-hbic/tree/rsfMRI-qc-paper>.

Ethics statement

All procedures involving human participants were performed in accordance with the ethical standards of the Declaration of Helsinki, and the study was approved by the Institutional Review Board where the data were collected. Informed consent was obtained from all participants.

Author contributions

RL, H-WY, BM, WB, JH, and LM conceived and designed the approach. RL, H-WY, BM, JH, RK, VP, and LM contributed to the analysis. All authors contributed to the interpretation of the data, drafting and revising of the manuscript, and provide approval for the manuscript.

Funding

This work was funded in part by the Hoglund Biomedical Imaging Center—which is supported by a generous gift from Forrest and Sally Hoglund—and funding from the National Institutes of Health: S10 RR29577 to the Hoglund Biomedical Imaging Center, UL1 TR000001 to the Frontiers: Heartland Institute for Clinical and Translational Research (CTSA), and P30 AG035982 to the University of Kansas Alzheimer's Disease Research Center (KU ADRC).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

References

- Andellini, M., Cannata, V., Gazzellini, S., Bernardi, B., and Napolitano, A. (2015). Test-retest reliability of graph metrics of resting state MRI functional brain networks: a review. *J. Neurosci. Methods* 253, 183–192. doi: 10.1016/j.jneumeth.2015.05.020
- Beall, E. B., and Lowe, M. J. (2014). SimPACE: generating simulated motion corrupted BOLD data with synthetic-navigated acquisition for the development and evaluation of SLOMOCO: a new, highly effective slice-wise motion correction. *NeuroImage* 101, 21–34. doi: 10.1016/j.neuroimage.2014.06.038
- Birn, R. M., Molloy, E. K., Patriat, R., Parker, T., Meier, T. B., Kirk, G. R., et al. (2013). The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage* 83, 550–558. doi: 10.1016/j.neuroimage.2013.05.099
- Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Carp, J. (2013). Optimizing the order of operations for movement scrubbing: comment on Power et al. *NeuroImage* 76, 436–438. doi: 10.1016/j.neuroimage.2011.12.061
- Chen, S., Ross, T. J., Chuang, K. S., Stein, E. A., Yang, Y., and Zhan, W. (2010). A new approach to estimating the signal dimension of concatenated resting-state functional MRI data sets. *Magn. Reson. Imaging* 28, 1344–1352. doi: 10.1016/j.mri.2010.04.002
- Couvry-Duchesse, B., Blokland, G. A. M., Hickie, I. B., Thompson, P. M., Martin, N. G., de Zubicaray, G. I., et al. (2014). Heritability of head motion during resting state functional MRI in 462 healthy twins. *NeuroImage* 102, 424–434. doi: 10.1016/j.neuroimage.2014.08.010
- Couvry-Duchesse, B., Ebejer, J. L., Gillespie, N. A., Duffy, D. L., Hickie, I. B., Thompson, P. M., et al. (2016). Head motion and inattention/hyperactivity share common genetic influences: implications for fMRI studies of ADHD. *PLoS One* 11:e0146271. doi: 10.1371/journal.pone.0146271
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014
- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* 12:e0184661. doi: 10.1371/journal.pone.0184661
- Fox, M. D., and Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Front. Syst. Neurosci.* 4:19. doi: 10.3389/fnsys.2010.00019
- Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., et al. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage* 95, 232–247. doi: 10.1016/j.neuroimage.2014.03.034
- Herbst, M., Zahneisen, B., Knowles, B., Zaitsev, M., and Ernst, T. (2015). Prospective motion correction of segmented diffusion weighted EPI. *Magn. Reson. Med.* 74, 1675–1681. doi: 10.1002/mrm.25547
- Johnstone, T., Ores Walsh, K. S., Greischar, L. L., Alexander, A. L., Fox, A. S., Davidson, R. J., et al. (2006). Motion correction and the use of motion covariates in multiple-subject fMRI analysis. *Hum. Brain Mapp.* 27, 779–788. doi: 10.1002/hbm.20219
- Kong, X. Z., Zhen, Z., Li, X., Lu, H. H., Wang, R., Liu, L., et al. (2014). Individual differences in impulsivity predict head motion during magnetic resonance imaging. *PLoS One* 9:e104989. doi: 10.1371/journal.pone.0104989
- Lanka, P., and Deshpande, G. (2019). Combining prospective acquisition CorrEction (PACE) with retrospective correction to reduce motion artifacts in resting state fMRI data. *Brain Behav.* 9:e01341. doi: 10.1002/brb3.1341
- Lazar, N. A. (2008). "The statistical analysis of functional MRI data" in *Statistics for biology and health*. ed. M. Gail (New York, NY: Springer), 299.
- Lee, C. C., Grimm, R. C., Manduca, A., Felmlee, J. P., Ehman, R. L., Riederer, S. J., et al. (1998). A prospective approach to correct for inter-image head rotation in fMRI. *Magn. Reson. Med.* 39, 234–243. doi: 10.1002/mrm.1910390210
- Lepping, R. J., Bruce, A. S., Francisco, A., Yeh, H. W., Martin, L. E., Powell, J. N., et al. (2015). Resting-state brain connectivity after surgical and behavioral weight loss. *Obesity (Silver Spring)* 23, 1422–1428. doi: 10.1002/oby.21119
- Lepping, R. J., Honea, R. A., Martin, L. E., Liao, K., Choi, I. Y., Lee, P., et al. (2019). Long-chain polyunsaturated fatty acid supplementation in the first year of life affects brain function, structure, and metabolism at age nine years. *Dev. Psychobiol.* 61, 5–16. doi: 10.1002/dev.21780
- Little, R. J. A., and Rubin, D. B. (eds). (2002). "Statistical analysis with missing data" in *Wiley Series in probability and statistics*. 2nd ed (New York, NY: Wiley), 381.
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *Elife* 10:10. doi: 10.7554/eLife.71774
- Maziero, D., Rondinoni, C., Marins, T., Stenger, V. A., and Ernst, T. (2020). Prospective motion correction of fMRI: improving the quality of resting state data affected by large head motion. *NeuroImage* 212:116594. doi: 10.1016/j.neuroimage.2020.116594
- Muraskin, J., Ooi, M. B., Goldman, R. I., Krueger, S., Thomas, W. J., Sajda, P., et al. (2013). Prospective active marker motion correction improves statistical power in BOLD fMRI. *NeuroImage* 68, 154–161. doi: 10.1016/j.neuroimage.2012.11.052
- Pardoe, H. R., Kucharsky Hiess, R., and Kuzniecky, R. (2016). Motion and morphometry in clinical and nonclinical populations. *NeuroImage* 135, 177–185. doi: 10.1016/j.neuroimage.2016.05.005
- Patriat, R., Molloy, E. K., and Birn, R. M. (2015). Using edge voxel information to improve motion regression for rs-fMRI connectivity studies. *Brain Connect* 5, 582–595. doi: 10.1089/brain.2014.0321
- Patriat, R., Reynolds, R. C., and Birn, R. M. (2016). An improved model of motion-related signal changes in fMRI. *NeuroImage* 144, 74–82. doi: 10.1016/j.neuroimage.2016.08.051
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2013). Steps toward optimizing motion artifact removal in functional connectivity MRI; a reply to Carp. *NeuroImage* 76, 439–441. doi: 10.1016/j.neuroimage.2012.03.017
- Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* 84, 320–341. doi: 10.1016/j.neuroimage.2013.08.048
- Power, J. D., Schlaggar, B. L., and Petersen, S. E. (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. *NeuroImage* 105, 536–551. doi: 10.1016/j.neuroimage.2014.10.044
- Pruim, R. H., Mennes, M., Buitelaar, J. K., and Beckmann, C. F. (2015). Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. *NeuroImage* 112, 278–287. doi: 10.1016/j.neuroimage.2015.02.063
- Rogers, B. P., Morgan, V. L., Newton, A. T., and Gore, J. C. (2007). Assessing functional connectivity in the human brain by fMRI. *Magn. Reson. Imaging* 25, 1347–1357. doi: 10.1016/j.mri.2007.03.007
- Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughhead, J., Calkins, M. E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage* 64, 240–256. doi: 10.1016/j.neuroimage.2012.08.052
- Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., et al. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Hum. Brain Mapp.* 35, 1981–1996. doi: 10.1002/hbm.22307
- Thesen, S., Heid, O., Mueller, E., and Schad, L. R. (2000). Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magn. Reson. Med.* 44, 457–465. doi: 10.1002/1522-2594(200009)44:3<457::AID-MRM17>3.0.CO;2-R
- Thulborn, K. R. (1999). Visual feedback to stabilize head position for fMRI. *Magn. Reson. Med.* 41, 1039–1043. doi: 10.1002/(SICI)1522-2594(199905)41:5<1039::AID-MRM24>3.0.CO;2-N
- Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage* 59, 431–438. doi: 10.1016/j.neuroimage.2011.07.044
- Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., and Castellanos, F. X. (2015). Inscapes: a movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage* 122, 222–232. doi: 10.1016/j.neuroimage.2015.07.069

Frontiers in Neuroscience

Provides a holistic understanding of brain
function from genes to behavior

Part of the most cited neuroscience journal series
which explores the brain - from the new eras
of causation and anatomical neurosciences to
neuroeconomics and neuroenergetics.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

