A multi-dataset evaluation of frame censoring for task-based fMRI

Michael S. Jones, Zhenchen Zhu*, Aahana Bajracharya, Austin Luor, Jonathan E. Peelle

Department of Otolaryngology, Washington University in St. Louis, St. Louis MO USA
* Current affiliation: MD Program, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing China

Running title: Motion correction in task-based fMRI

Keywords: motion correction, head movement, frame censoring, scrubbing, FD, DVARS, task-based fMRI

Please address correspondence to:

Dr. Michael Jones
Department of Otolaryngology
Washington University in St. Louis
660 South Euclid, Box 8115
St. Louis, MO 63110
email: jones.mike@wustl.edu

Dr. Jonathan Peelle
Department of Otolaryngology
Washington University in St. Louis
660 South Euclid, Box 8115
St. Louis, MO 63110
email: jpeelle@wustl.edu

Abstract

Subject motion during fMRI can affect our ability to accurately measure signals of interest. In recent years, frame censoring—that is, statistically excluding motion-contaminated data within the general linear model using nuisance regressors—has appeared in several task-based fMRI studies as a mitigation strategy. However, there have been few systematic investigations quantifying its efficacy. In the present study, we compared the performance of frame censoring to several other common motion correction approaches for task-based fMRI using open data and reproducible workflows. We analyzed eight datasets available on OpenNeuro.org representing eleven distinct tasks in child, adolescent, and adult participants. Performance was quantified using maximum t-values in group analyses, and ROI-based mean activation and split-half reliability in single subjects. We compared frame censoring to the use of 6 and 24 canonical motion regressors, wavelet despiking, robust weighted least squares, and untrained ICA-based denoising. Thresholds used to identify censored frames were based on both motion estimates (FD) and image intensity changes (DVARS). Relative to standard motion regressors, we found consistent improvements for modest amounts of frame censoring (e.g., 1–2% data loss), although these gains were frequently comparable to what could be achieved using other techniques. Importantly, no single approach consistently outperformed the others across all datasets and tasks. These findings suggest that although frame censoring can improve results, the choice of a motion mitigation strategy depends on the dataset and the outcome metric of interest.

Introduction

Obtaining high-quality neuroimaging data depends on minimizing artifacts. Although advancements in hardware and pulse sequence design have reduced many types of noise inherent to functional MRI, other sources remain (Bianciardi et al., 2009). One prominent challenge is artifacts caused by subject head motion. Among other effects, head motion changes the part of the brain sampled by a particular voxel and can introduce changes in signal intensity through interactions with the magnetic field, which add noise to the data and make it harder to identify signals of interest.

The effects of head motion have received recent scrutiny in the context of resting state functional connectivity. Because motion-related artifacts occur in many voxels simultaneously, they can introduce correlations in fMRI time series that are unrelated to BOLD activity, leading to inaccurate estimates of functional connectivity (Power et al., 2015; Satterthwaite et al., 2019). However, spurious activation is also of concern in task-based functional neuroimaging. Rigid body realignment—a mainstay of fMRI analysis for decades—goes some way towards improving correspondence across images (Ashburner and Friston, 2004), but does not remove extraneous signal components introduced by movement (Friston et al., 1996). A common approach for mitigating motion-related artifacts is to include the 6 realignment parameters (translation and rotation around the X, Y, and Z axes) as nuisance regressors in first-level models.

Alternatively, several data-driven strategies have been developed to reduce the influence of high-motion scans on estimated activations. Wavelet decomposition identifies artifacts by exploiting their non-stationarity across different temporal scales (Patel et al., 2014). The method has been applied in resting state studies but is also applicable to task-based data. Independent component analysis (Pruim et al., 2015) identifies artifacts based on the spatial distribution of shared variance. In robust weighted least squares (Diedrichsen and Shadmehr, 2005), a two-pass modeling procedure is used to produce a collection of nuisance regressors which are then included in the final analysis to weight frames by the inverse of their variance (that is, downweighting frames with high error).

An alternative motion correction strategy is “scrubbing” or “frame censoring” (Lemieux et al., 2007; Siegel et al., 2014). In this approach, bad scans are identified and excluded from statistical analysis. One approach is to do so by modeling them in the general linear model using nuisance regressors (i.e. “scan-nulling regressors” or “one-hot encoding”). Although frame censoring has received considerable interest in resting state fMRI over the past several years (Power et al., 2012; Gratton et al., 2020a), it has not seen widespread use in the task-based fMRI literature. Censoring approaches involve some effective data loss, in that censored frames do not contribute to the task-related parameter estimates, and that columns introduced to the design matrix to perform censoring reduce the available degrees of freedom. Choosing an appropriate metric and associated threshold for identifying bad scans can also be challenging. Thus, additional information over what threshold should be used for identifying bad frames—and relatedly, how much data is lost vs. retained—is necessary to make informed decisions.

Although several published studies comparing differing correction strategies exist (Ardekani et al., 2001; Oakes et al., 2005; Johnstone et al., 2006), a drawback of prior work is that evaluation was often limited to a single dataset (see **Supplemental Table 1**). The degree to which an optimal strategy for one dataset generalizes to other acquisition schemes, tasks, or populations is not clear. With the increased public availability of neuroimaging datasets (Poldrack et al., 2013; Markiewicz et al., 2021), the possibility of evaluating motion correction approaches across a range of data has become more feasible.

In the present work, we sought to compare the performance of identical pipelines on a diverse selection of tasks, using data from different sites, scanners, and subject pools.

Although our primary interest was frame censoring, we considered seven different motion-correction approaches:

1. six canonical head motion estimates (RP6)
2. 24-term expansions of head motion estimates (RP24)
3. wavelet despiking (WDS)
4. robust weighted least squares (rWLS)
5. untrained independent component analysis (ICA)
6. frame censoring based on frame displacement (FD)
7. frame censoring based on variance differentiation (DVARS)

This list is not exhaustive but representative of approaches that are currently used and feasible to include in an automated processing pipeline.

Because it is impossible to determine a "ground truth" result with which to compare the effectiveness of these approaches, we instead considered three complementary outcome metrics: 1) the maximum group t-statistic both across the whole-brain and in a region-of-interest relevant to the task; 2) the average parameter estimates from within the same ROI (that is, effect size); and 3) the degree of test-retest consistency exhibited by subject-level parametric maps. These metrics are simple to define yet functionally meaningful, and can be applied to data from almost any fMRI study.

Methods

Datasets

We analyzed eight studies obtained from OpenNeuro (Markiewicz et al., 2021), several of which included multiple tasks or multiple participant groups. As such, the eight selected studies provided a total of 15 datasets. The selection process was informal, but studies given priority included 1) a clearly-defined task, 2) a sufficient number of subjects to allow second-level modeling, 3) sufficient data to make test-retest evaluation possible, and 4) a publication associated with the data describing a result to which we could compare our own analysis.

A summary of the eight datasets selected is shown in **Table 1** (acquisition details provided in **Supplemental Table 2**). Additional information, including task details, modeling/contrast descriptions compiled from publication(s) associated with a given study, and any data irregularities encountered during analysis, is provided in the **Supplemental Materials**.

Analysis

Analysis was performed using Automatic Analysis version 5.4.0 (Cusack et al., 2015) (RRID: SCR_003560), which scripted a combination of SPM12 (Wellcome Trust Centre for Neuroimaging) version 7487 (RRID: SCR_007037) and FMRIB Software Library (FSL; FMRIB Analysis Group; (Jenkinson et al., 2012) version 6.0.1 (RRID: SCR_002823). BrainWavelet Toolbox v2.0 (Patel et al., 2014) was used for wavelet despiking, and rWLS version 4.0 (Diedrichsen and Shadmehr, 2005) for robust weighted least squares. Analysis scripts used in the study are available at https://osf.io/n5v3w/.

To the extent possible, we used the same preprocessing pipeline for all datasets (**Figure 1a**). Briefly, structural and functional images were translated to the center of the scanned volume and the first four frames of each session were removed in functional images to allow for signal stabilization. This was followed by bias correction of the structural image, realignment, coregistration of the functional and structural images, normalization into MNI space using a unified segmentation approach (Ashburner and Friston, 2005) resampled at 2 mm, and smoothing of the functional images using an 8 mm FWHM Gaussian kernel.

Table 1. Summary of datasets analyzed

Dataset	Reference	Task	Age group*	# subs	FD (median ± SD)	frames per subject
ds000102	Kelly et al. (2008)	flanker	YA	22	0.11 ± 0.12	284
ds000107	Duncan et al. (2009)	1-back	YA	43	0.08 ± 0.14	323
ds000114	Gorgolewski et al. (2013a)	motor (lips)	YA	10	0.14 ± 0.16	360
		covert verb	YA	10	0.11 ± 0.11	338
		overt word	YA	10	0.13 ± 0.12	144
		line bisection	YA	9	0.13 ± 0.18	468
ds000228	Richardson et al. (2018)	movie viewing	C	122	0.21 ± 0.93	164
			YA	33	0.18 ± 0.27	164
ds001497	Lewis-Peacock and Postle (2008)	face perception	YA	10	0.11 ± 0.12	1146
ds001534	Courtney et al. (2018)	food images	YA	42	0.10 ± 0.16	552
ds001748	Fynes-Clinton et al. (2019)	memory retrieval	C	21	0.16 ± 0.36	438
			T	20	0.12 ± 0.17	438
			YA	21	0.08 ± 0.17	438
ds002382	Rogers et al. (2020)	word recognition	YA	29	0.14 ± 0.35	710
			OA	32	0.30 ± 0.34	710

Note: * OA = older adults; YA = young adults; T = teens; C = children

Functional images were corrected for motion artifacts using each of the following approaches: 1) inclusion of six canonical motion estimates in the first-level model as nuisance regressors, 2) inclusion of 24 nuisance regressors based on a second-order expansion of the motion estimates and first derivatives, 3) wavelet despiking, 4) robust weighted least squares, 5) ICA denoising, 6) frame censoring based on framewise displacement (FD) or 7) differential variance (DVARS) thresholding (FD/DVARS thresholding is described below).

Statistical modeling was performed in SPM in all motion correction approaches. First-level modeling included a contrast of interest described in a publication associated with the dataset for evaluation, followed by second-level analysis to produce group-level statistical maps. All first- and second-level t-maps were thresholded at a voxelwise threshold of $p < 0.001$ (uncorrected).

Minor pipeline modifications were required for robust weighted least squares, wavelet despiking, and ICA denoising. As recommended by developers of the rWLS toolbox, unsmoothed data was used for variance estimation and contrast maps were smoothed after modeling. For wavelet despiking, functional images were rescaled to a whole-brain median of 1000 across all frames prior to processing. The default toolbox settings (wavelet: d4, threshold: 10, boundary: reflection, chain search: moderate, scale number: liberal) were used. Finally, ICA-based denoising was implemented using ICA-AROMA (Pruim et al., 2015) with additional processing steps performed within FSL. Briefly, the unsmoothed coregistered functional image was demeaned, detrended, smoothed, and then nonlinearly warped to the FSL 2 mm MNI152 template using FNIRT. The normalized functional image was then passed to AROMA for denoising.

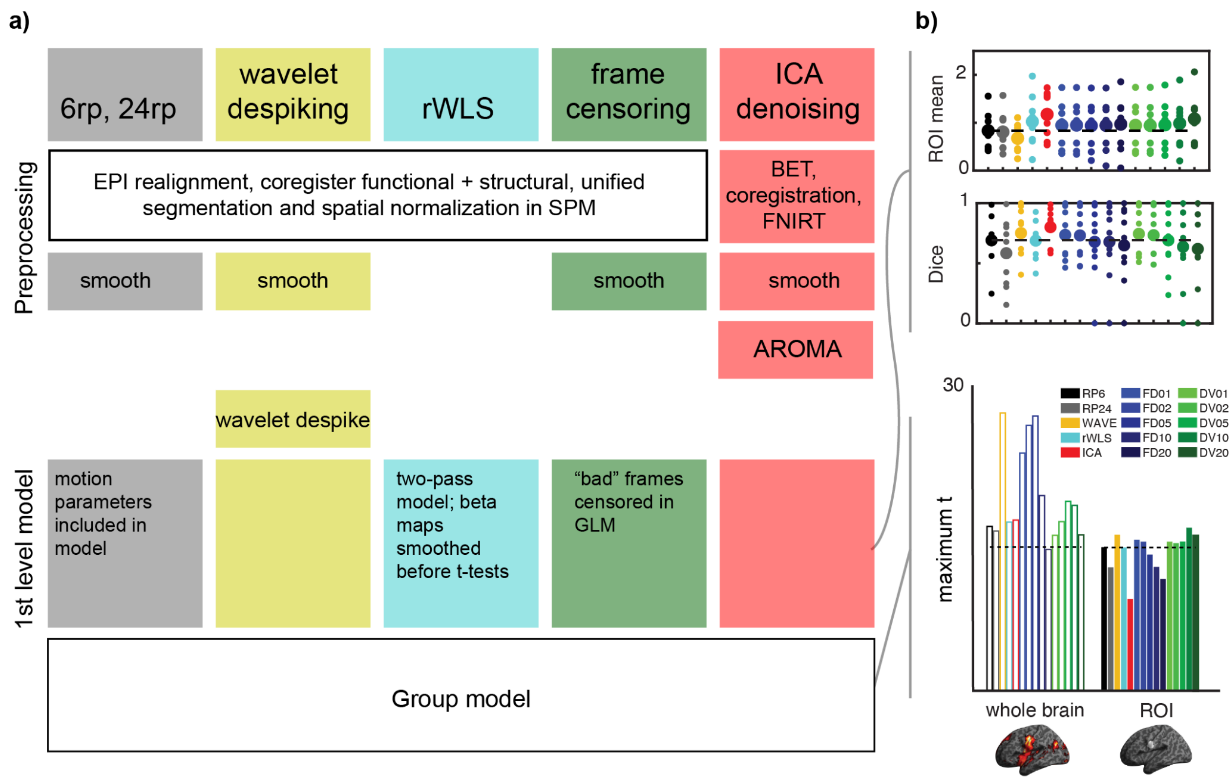


Figure 1. Schematic of processing pipeline and outcome measures. **(a)** Summary of preprocessing and model steps in common and differing across motion correction strategies. **(b)** Following statistical modeling, outcomes are summarized in mean parameter estimates and Dice overlap of thresholded single-subject maps (top) and maximum t-value from the group analysis (bottom).

Evaluation of Motion Correction Performance

Three measures were used to quantify the performance of each motion correction strategy, illustrated in **Figure 1b**: 1) maximum t-value, 2) effect size, and 3) subject replicability. In the first measure, the maximum t-value occurring in the group level parametric map was extracted both at the whole-brain level and also within a region-of-interest relevant to the task. The effect size was quantified as the mean of all voxels within the ROI for each subject using the first-level beta maps. To evaluate subject replicability, session data were treated as a test-retest paradigm (the first session versus the second session in studies having fewer than three sessions; even-numbered versus odd-numbered sessions otherwise). Replicability was quantified as the Dice coefficient of thresholded first-level t-maps (0.001, uncorrected) in each subject (restricted to the ROI).

FD and DVARS Thresholding

Motion correction approaches based on frame censoring required quantification of motion artifacts which could then be subjected to thresholding. Both framewise displacement (FD) and differential variance (DVARS) were used. Framewise displacement was calculated as the sum of the six head motion estimates obtained from realignment, with a dimensional conversion of the three rotations assuming the head is a 50 mm sphere (Power et al., 2012). DVARS was calculated as the root-mean-squared of the time difference in the BOLD signal calculated across the entire brain (Smyser et al., 2011). As shown in **Figure 2A**, both metrics closely tracked artifacts apparent in voxel intensities and also each other. Although FD and DVARS in a given session tended to be correlated (**Figure 2B**), they were not identical and could exhibit slightly different time courses and relative peak amplitudes. As such, we explored the use of both measures.

Thresholds were determined by calculating FD and DVARS across all sessions in all subjects, which allowed values to be identified that resulted in 1%, 2%, 5%, 10%, and 20% frame violations across the entire dataset (**Figure 2C**). We adopted this strategy rather than using a fixed value of FD or DVARS for several reasons. First, FD and DVARS magnitudes change with the TR of the data, because the TR is the sampling rate (for a given movement, sampling more rapidly will give smaller FD values, even though the total motion is the same). Secondly, different calculations of FD provide different values (Jenkinson et al., 2002; Power et al., 2012; Van Dijk et al., 2012), and thus any absolute threshold would necessarily be metric-specific. Finally, datasets differ in their tasks and populations, and we anticipated that a constant threshold would not be suitable for all datasets. We, therefore, employed the frame-percent thresholding strategy in order to obtain an informative range of results in all studies examined. Because the threshold is chosen to limit data loss in the whole group, it allows high-motion subjects to have more frames censored than low-motion subjects, which was one of our primary goals.

The threshold values that resulted from percent data loss targeting in these datasets are shown in **Supplemental Figure 1**.

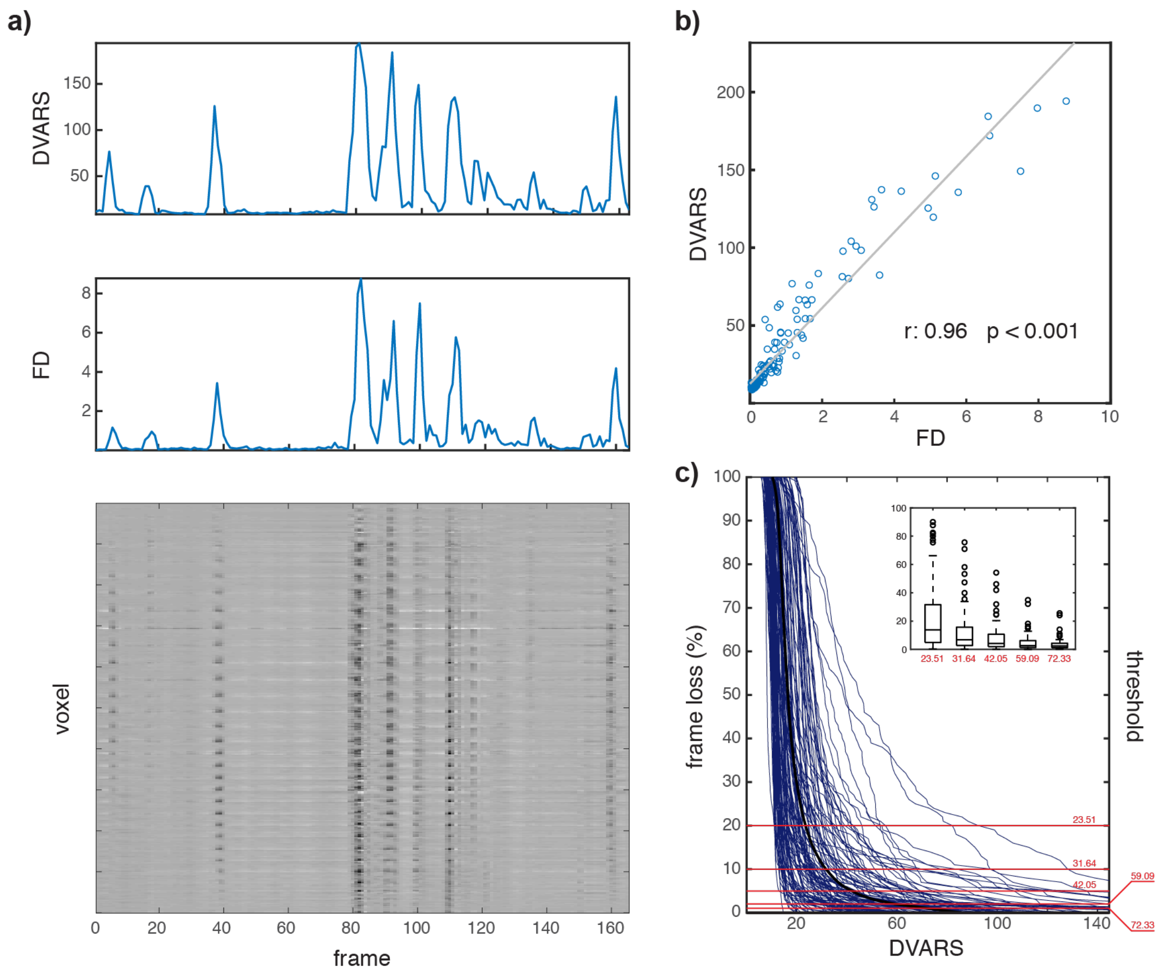


Figure 2. Calculation of frame censoring thresholds. **(a)** Example grayplot (Power, 2017) showing 500 random gray matter voxels (single subject data from ds001534). DVARS and FD for this session are plotted above. Spikes in the metrics can be used to identify frames contaminated by artifacts. **(b)** DVARS and FD are correlated but exhibit differing amplitudes and time courses. As such, the use of both was explored. **(c)** The collection of metric values (here shown for DVARS) used for frame censoring was determined by plotting frameloss for each subject as a function of threshold (thin blue traces). Interpolation of the mean response (thick black trace) then provided estimates of metric values corresponding to a target data loss of 1%, 2%, 5%, 10%, or 20%. Box plot (inset) summarizes the resulting data loss across all subjects at each threshold (*box*: 25-75% percentiles; *crosses*: >3 SD outliers).

To impose frame censoring, first-level modeling was repeated for each threshold with a delta function (i.e. a scan-nulling regressor) included in the design matrix at the location of each violation, which effectively removes the contribution of the targeted frame from the analysis.

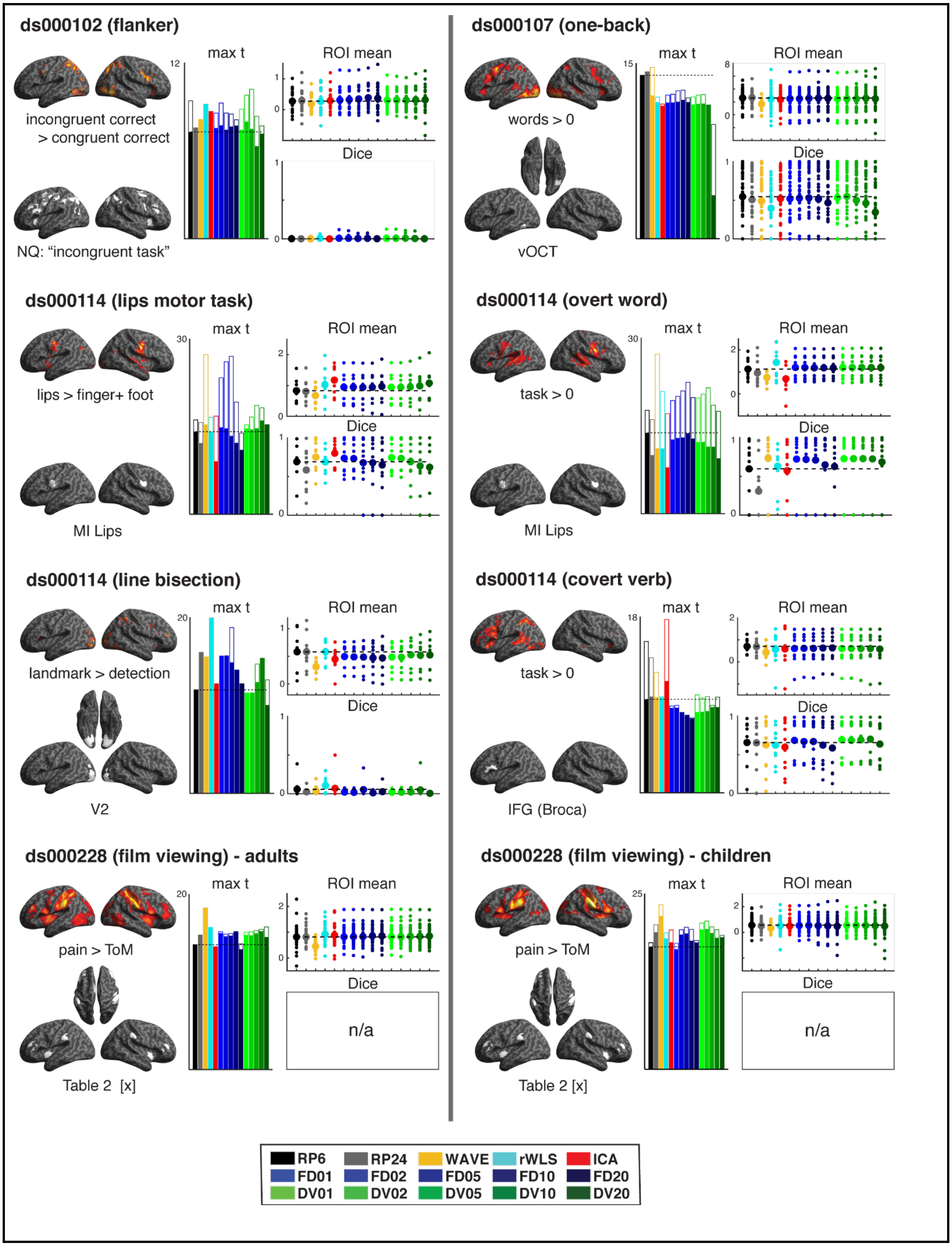
Region of Interest Definition

A task-relevant ROI for each study/task was defined in one of three ways: 1) a 5-mm sphere (or spheres) centered at coordinates reported in a publication associated with the dataset, 2) a whole-brain Z-mask generated by a task-relevant search term (e.g., "incongruent task") in NeuroQuery (Dockès et al., 2020) and thresholded z > 3, or 3) a binarized tissue probability map in the SPM Anatomy Toolbox (Eickhoff et al., 2005) for a task-relevant brain structure or anatomical region (e.g., "V2").

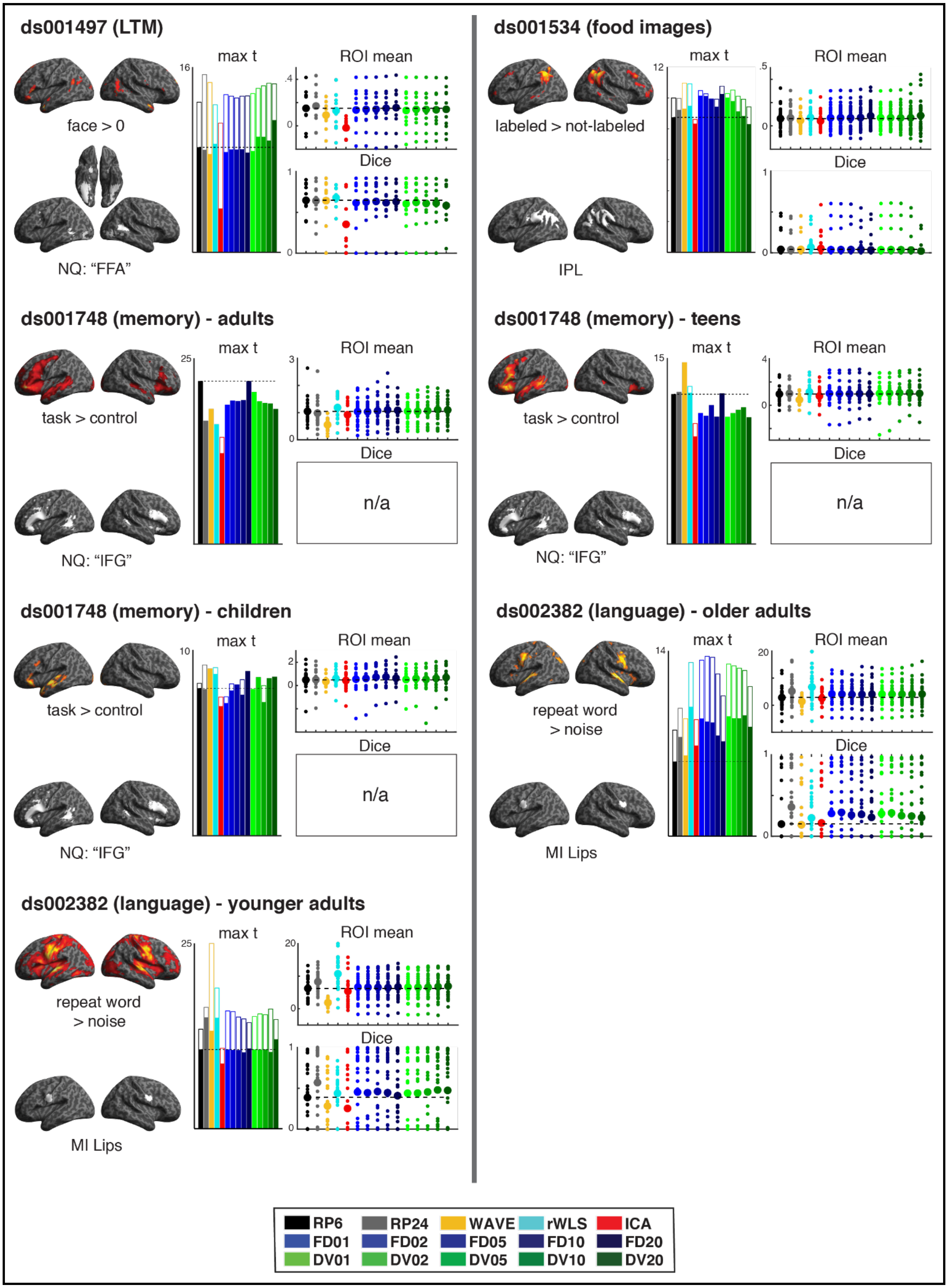
Results

Performance of the motion correction strategies organized by dataset is shown in **Figure 3**. Each panel includes a representative second-level thresholded t-map at the upper left ($p < 0.001$, uncorrected) using the "RP6" approach (six canonical motion parameters included as nuisance regressors). A contrast descriptor is given below the map. The ROI used for evaluation is shown at lower left with the source listed under the rendered image; "NQ" indicates search term from NeuroQuery (Dockès et al., 2020); all other labels indicate either an Anatomy Toolbox tissue probability map (Eickhoff et al., 2005) or a 5 mm sphere. Additional details on ROI definition used in each analysis are provided in the **Supplemental Materials**.

These results show there is substantial variability in motion correction approaches, with performance depending both on the data under consideration and the chosen performance metric. However, some general trends are apparent. Wavelet despiking tended to offer the best maximum t-value in both the whole-brain and ROI-constrained evaluation, with robust weighted least squares also exhibiting good performance (note the ROI-restricted maximum t-value, shown in filled bars, are superimposed on the whole-brain results, shown in open bars in Figure 3 due to space restrictions). Conversely, ICA gave consistently poorer results although it offered the best maximum t-value in the ds000114 covert verb task. Performance of FD and DVARS frame censoring were highly variable, with the application of increasingly stringent thresholds improving performance in some datasets while decreasing it in others. A somewhat consistent behavior is a loss of performance at the highest (20%) FD or DVARS threshold. As a rule, frame censoring performed better than RP6 and RP24 motion correction, although RP6 is competitive (if not optimal) in both ds000107 and ds001748.


ds000102 (flanker)
incongruent correct > congruent correct
NQ: "incongruent task"
ds000107 (one-back)
words > 0
vOCT
ds000114 (lips motor task)
lips > finger+ foot
MI Lips
ds000114 (overt word)
task > 0
MI Lips
ds000114 (line bisection)
landmark > detection
V2
ds000114 (covert verb)
task > 0
IFG (Broca)
ds000228 (film viewing) - adults
pain > ToM
Table 2 [x]
ds000228 (film viewing) - children
pain > ToM
Table 2 [x]
max t
ROI mean
Dice
n/a
RP6
RP24
WAVE
rWLS
ICA
FD01
FD02
FD05
FD10
FD20
DV01
DV02
DV05
DV10
DV20

262
263



ds001497 (LTM)
face > 0
NQ: "FFA"
max t
ROI mean
Dice
ds001534 (food images)
labeled > not-labeled
IPL
ds001748 (memory) - adults
task > control
NQ: "IFG"
n/a
ds001748 (memory) - teens
task > control
NQ: "IFG"
ds001748 (memory) - children
task > control
NQ: "IFG"
ds002382 (language) - older adults
repeat word > noise
MI Lips
ds002382 (language) - younger adults
repeat word > noise
MI Lips
RP6 RP24 WAVE rWLS ICA
FD01 FD02 FD05 FD10 FD20
DV01 DV02 DV05 DV10 DV20

Figure 3. Summary of denoising algorithm performance for all datasets examined in the study. Each panel includes a representative thresholded group t-map at left (p=0.001, uncorrected) for the given contrast with the ROI used for evaluation plotted below. At center, ROI-restricted maximum t-values are superimposed on whole-brain results for each denoising approach. Plots at right show individual-subject mean ROI effect size (*top*) and Dice coefficient for a split-half test-retest evaluation (*bottom*). Datasets that did not permit test-retest evaluation are noted "n/a." The horizontal dotted line in a given plot indicates RP6 results for reference.

The mean effect size shown in these results is largely insensitive to the selected motion correction approach. The two exceptions are wavelet despiking and ICA, which produced consistently smaller values than the other approaches. This may reflect suboptimal parameter selection in these algorithms (see Discussion). Robust weighted least squares offered competitive results in all datasets and notably superior results in ds002382 and the ds000114 overt word task. FD and DVARS frame censoring neither improved nor degraded results regardless of threshold, producing a mean effect size indistinguishable from both the RP6 and RP24 approaches save for a few isolated individual subjects.

The test-retest results also demonstrate a great deal of variability. The Dice coefficients exhibit substantial inter-subject differences, resulting in a mean performance that is similar across all motion correction strategies. However, excluding ds000102, ds001534, and the ds000114 line bisection task which provide an uninformative test-retest quantification, some trends can be identified. There is a detectable decrease in both the FD and DVARS frame censoring results, especially at 20% thresholding. In general, all differences are minor, save for ICA which performs notably better in the ds000114 motor task and notably worse in ds001487.

A summary of these results is shown in **Figure 4a**, in which average values of the four performance metrics are plotted for all 15 datasets/tasks. Several of the trends noted above remain apparent. Wavelet despiking gives the largest whole-brain maximum t-value. Robust weighted least squares resulted in the best ROI-constrained performance. Light-to-moderate frame censoring results in improvement which then declines as more aggressive thresholding is applied. Robust weighted least squares produces the largest average effect size. Wavelet despiking and ICA produce poor results as measured by this metric. Finally, the averaged Dice coefficient is less than 0.5 in all datasets. A decline of FD and DVARS frame censoring performance with increasing threshold is apparent. However, all of the test-retest results exhibit substantial variability (error bars denote +1 SD in the maximum t-value plot; +/- 1 SD in ROI mean effect size and Dice).

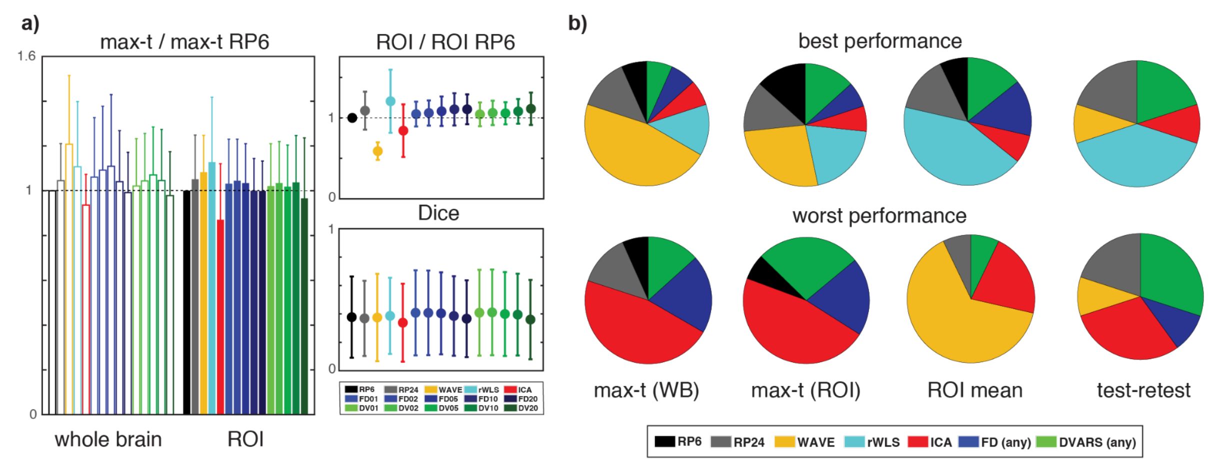


Figure 4. (a) Whole-brain and ROI-restricted maximum t-values (*left*), mean effect size (*upper right)* and test-retest Dice coefficient (*lower right)* averaged across all datasets. A value of 1 indicates equivalent performance to the "standard" RP6 approach. **(b)** Motion correction performance summarized as the proportion of datasets a given approach gave the best (top row) or worst (bottom) results as measured by the four performance metrics.

An alternate summary of algorithm performance is presented in **Figure 4b**, in which the best and worst performer measured by each metric was identified in each of the 15 datasets and the resulting proportions plotted as pie charts. The trends and variability evident in the grand averages are also apparent in these data. Robust weighted least squares offered best performance on many datasets and worst performance on none. Wavelet despiking gave the best maximum t-value in about half (whole-brain) or one quarter (ROI-restricted) of the studies, but the worst ROI mean effect size in over half. ICA denoising was often the worst performer, yet gave best results across all four metrics in at least one dataset. Frame censoring performed roughly equally well (or equally poorly) using either FD or DVARS, with one notable result being FD was never the worst performer for effect size. Finally, the performance of the RP6 and RP24 approaches are middling, producing best or worse maximum t-value on only one or two datasets and, with one exception, never producing best nor worst ROI mean or test-retest results.

Discussion

We explored the performance of a variety of approaches to correcting motion-related artifacts in task-based fMRI. The studies examined represent a broad range of task domains, including sensory, motor, language, memory, and other cognitive functions, with participants varying in age, sex, and other characteristics. Although we set out expecting to find converging evidence for an optimal strategy, instead our results demonstrate that the performance of motion correction approaches depends on both the data and the outcome of interest. We review our selected metrics below—whole-brain and ROI-restricted maximum t-value, mean effect size, and test-retest repeatability—followed by some general comments on each motion correction approach.

Comparing outcome metrics

The use of whole-brain maximum t-value measured in group-level statistical maps has the advantage that it requires few assumptions about the data or the expected pattern of activity. However, we did not observe a consistent pattern regarding which motion correction approach optimized the whole-brain maximum t-value. Disparity was even evident between different participant groups within a given study. For example, wavelet despiking had the highest whole-brain t statistic in ds0001748 in teens but RP6 offered better performance in adults.

In addition to whole-brain statistics, we examined maximum t-values within a selected region of interest. Our rationale for doing so was that researchers interested in task-based effects frequently have prior intuitions about where the most informative results are localized. A potential downside of this approach is the need to specify an ROI to examine. We found that motion correction approaches can exhibit substantially different whole-brain and ROI-restricted performance. In the ds000114 overt word task, for example, RP6 offered best performance within the motor cortex but poor performance in a whole-brain evaluation. Furthermore, frame censoring performance improved in some datasets but degraded in others as more stringent thresholding was applied. Obviously, a challenge inherent in such an evaluation is the actual ROI selection. Although we believe our choices are sensible, selection of a different ROI set may well result in a different overall view of performance.

To complement these group-level measures, we also considered two single-subject metrics: mean effect size and test-retest repeatability measured by Dice overlap in thresholded t maps. Effect size permits an examination of parameter estimates, and our use of averaging offers a direct and simple quantification. However, with the exceptions of wavelet despiking and aggressive frame censoring (revisited below), we observed that effect size was largely insensitive to the choice of motion correction strategy, although less than the variability observed in maximum t-value. This suggests the main effect of different motion correction approaches is a differential reduction in model error variance. If parameter estimation is the primary result of interest, then the choice of motion correction strategy may not be critical.

The test-retest evaluation was perhaps the least helpful result, with the performance of all motion correction approaches essentially indistinguishable under this metric. Although the outcome is disappointing, it should be noted that many of the studies included here were not designed to include a split-half repeatability analysis. It may be that more data per subject may be needed for this metric to be informative. In that sense, our analyses speak to the general challenges of obtaining reliable single-subject data in fMRI (Smith et al., 2005; Bennett and Miller, 2010; Gorgolewski et al., 2013b; Elliott et al., 2020), at least under conventional scanning protocols (Gratton et al., 2020b). Investigators of resting state fMRI have confronted a similar issue, and recommendations have appeared in the resting state fMRI literature outlining the minimal scan time required for reproducible results (Birn et al., 2013; Laumann et al., 2015). Perhaps an analogous standard might be possible for task-based fMRI, although any guideline would necessarily require the cognitive complexity of the task under investigation to be considered.

Comparing Motion Correction Approaches

No single denoising approach exhibited optimal performance on all datasets and all metrics. Algorithm performance did not appear to be systematically related to the nature of the task, acquisition parameters, nor any feature of the data that could be identified.

Interestingly, computationally-intensive approaches did not necessarily perform better than basic corrective measures. For some datasets, including six motion estimates as continuous nuisance regressors—a standard approach used in functional imaging for decades—could perform as well or better than more sophisticated algorithms that have emerged in recent years. Increasing the head motion estimate from a 6- to a 24-parameter expansion led to an improvement in some data but degraded results in others. Although such results are rather counterintuitive, we can provide a few observations, even if these data do not currently permit conclusive recommendations.

Two motion correction approaches that showed generally strong performance were wavelet despiking (WDS) and robust weighted least squares (rWLS). Together, these approaches offered best performance in approximately half of the datasets across all performance metrics (**Figure 4**). Additionally, in no evaluation did rWLS produce the worst results. In a statistical sense, robust weighted least squares might be seen as an optimal solution, in that it uses the error in the model to re-weight time points, reducing the influence of motion on parameter estimates. However, we also found that other motion correction strategies supplied similar, or superior, performance in several instances. One reason might be that rWLS linearly weights time points inversely related to their variance. To the degree that motion artifacts include a nonlinear component, linear weighting may not adequately (or at least, not optimally) remove all of the artifact.

In contrast to good performance of wavelet despiking as measured by maximum t-value, it gave notably low scores on mean effect size. However, this finding may simply reflect data scaling specific to the WDS implementation. It should also be noted the WDS toolbox offers 20 wavelets, and additional options that control algorithm behavior such as thresholding and chain search selection. The results obtained here are what can be expected using the default settings recommended by the toolbox developers, which includes a median 1000 rescaling of the functional data (and hence the lower effect size). Thus, numeric comparison to other approaches (that do not include rescaling) are problematic. It also may be possible to improve performance–including obtaining effect sizes concomitant with other motion correction approaches if that is judged critical by tuning the algorithm, although it is unclear how that process could be automated.

One unexpected result was the relatively poor performance of ICA denoising. Although individual exceptions exist, the approach produced consistently low scores on all evaluation metrics. However, ICA denoising was implemented here using FSL's ICA-AROMA. This package was selected because it does not require classifier training. More sophisticated ICA denoising tools such as MELODIC or ICA-FIX involve a visual review of training data to generate a set of noise classifiers based on the temporal, spatial, and frequency characteristics of identified artifacts (Salimi-Khorshidi et al., 2014; Griffanti et al., 2017). These options were not considered here because we sought to evaluate tools for motion correction that could be implemented in an automated pipeline. The potential of ICA for denoising task-based data should not be dismissed; rather, our results only indicate that the use of untrained ICA is probably suboptimal compared to other options, many of which are also less computationally intensive.

Frame censoring has appeared in several recent task-based studies (O'Hearn et al., 2016; Bakkour et al., 2017; Davis et al., 2017). In fact, it was an experience with frame censoring in the analysis of in-scanner speech production (Rogers et al., 2020) that motivated our interest in comparing motion correction approaches. We found that modest levels of frame censoring (e.g., 2–5% data loss) revealed a regional activation in high-motion subjects that appeared in low-motion subjects but was not apparent when standard (RP6) motion compensation was used. This suggested that use of a discrete rather than a continuous nuisance regressor may better preserve task variance in some applications. However, a more nuanced picture emerges from the present results, which suggest frame censoring is neither universally superior to nor worse than RP6. One possibility is that frame censoring performance involves a complex interaction between data quantity and quality. As each censored frame introduces an additional regressor to the design matrix, eventually the reduction in error variance may be overwhelmed by a loss of model degrees of freedom. This is anecdotally supported by a decline in many of the metric results observed here at the most stringent FD or

DVARS thresholds, an effect that was even more pronounced when 40% maximal censoring was explored in pilot work (data not shown). A prerequisite to improving frame censoring performance in future work would be to quantify this tradeoff.

One might argue that frame censoring should be based on a selected fixed threshold rather than a targeted percent data loss. The present results offer only mixed support for such a position. We investigated applying a fixed FD threshold of 0.9 to these data (**Supplemental Figure 1**). This value was used by Siegel and colleagues (2014) in their exploration of frame censoring and has since been used in published functional studies (e.g., Davis et al., 2017). In most of the datasets considered here, a 0.9 FD threshold would have resulted in less than 1% of frames being censored. This would be a reasonable amount of data loss, and might lead to some improvements compared to a standard RP6 approach (although we did not test this directly). However, ds000228/adults, ds001748/teens, and ds002382/YA would have incurred a 1-2% data loss, ds001748/child and ds002382/OA approximately 5% data loss, and ds000228/child approximately 13% data loss. These outcomes do not correspond to the best performance obtained across all approaches. Whole-brain or ROI-constrained maximum-t metrics peak at these values in some, but not all, datasets. Mean effect size and Dice coefficients add little to the evaluation as they appear largely insensitive to frame censoring thresholds in this range. Taken together, these results suggest there is no single threshold value that will optimize frame censoring for all applications.

Finally, it should be noted that we have focused on retrospective correction—that is, strategies for dealing with motion in existing data. A complementary approach would be to reduce head motion during acquisition. Protocols have been developed that offer promise to reduce subject motion, including movie viewing (Greene et al., 2018), custom head molds (Power et al., 2019), and providing feedback to participants (Dosenbach et al., 2017; Krause et al., 2019). However, these have not yet been widely adopted, nor are all compatible with task-based fMRI. With increasing awareness of the challenges caused by participant motion, perhaps greater interest in motion reduction (as opposed to mitigation) will follow.

Clearly, the present results do not identify unequivocal guidelines to select a motion correction strategy. Given the variability observed across datasets, with identical processing pipelines, exploring multiple strategies in a given dataset may be the best way of reducing motion artifacts, adding another set of parameters to an already large space of possible analyses (Carp, 2012; Poldrack et al., 2017; Botvinik-Nezer et al., 2020). Our results suggest that—frustratingly—no single motion correction strategy will give optimal results on every metric in every study, and that choices require considering both the nature of the specific data of interest and the most relevant outcome measure.

Acknowledgments

This work was supported by grants R01 DC014281, R01 DC016594, R21 DC016086, and T32 EB014855 (to A.B.) from the US National Institutes of Health. OpenNeuro is supported by NSF Grant OCI-1131441.

References

Ardekani BA, Bachman AH, Helpern JA (2001) A quantitative comparison of motion detection algorithms in fMRI. Magn Reson Imaging 19:959–963.

Ashburner J, Friston KJ (2004) Rigid Body Registration. In: Human Brain Function, Second. (Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Price CJ, Zeki S, Ashburner J, Penny W, eds), pp 635–653. New York: Elsevier.

Ashburner J, Friston KJ (2005) Unified segmentation. Neuroimage 26:839–851.

Bakkour A, Lewis-Peacock JA, Poldrack RA, Schonberg T (2017) Neural mechanisms of cue-approach training. Neuroimage 151:92–104.

Bennett CM, Miller MB (2010) How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci 1191:133–155.

Bianciardi M, Fukunaga M, van Gelderen P, Horovitz SG, de Zwart JA, Shmueli K, Duyn JH (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.

Birn RM, Molloy EK, Patriat R, Parker T, Meier TB, Kirk GR, Nair VA, Meyerand ME, Prabhakaran V (2013) The effect of scan length on the reliability of resting-state fMRI connectivity estimates. Neuroimage 83:550–558.

Botvinik-Nezer R et al. (2020) Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582:84–88.

Carp J (2012) On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front Neurosci 6:149.

Courtney AL, PeConga EK, Wagner DD, Rapuano KM (2018) Calorie information and dieting status modulate reward and control activation during the evaluation of food images. PLoS One 13:e0204744.

Cusack R, Vicente-Grabovetsky A, Mitchell DJ, Wild CJ, Auer T, Linke AC, Peelle JE (2015) Automatic analysis (aa): Efficient neuroimaging workflows and parallel processing using Matlab and XML. Front Neuroinform 8:90.

Davis T, Goldwater M, Giron J (2017) From Concrete Examples to Abstract Relations: The Rostrolateral Prefrontal Cortex Integrates Novel Examples into Relational Categories. Cereb Cortex 27:2652–2670.

Diedrichsen J, Shadmehr R (2005) Detecting and adjusting for artifacts in fMRI time series data. Neuroimage 27:624–634.

Dockès J, Poldrack RA, Primet R, Gözükan H, Yarkoni T, Suchanek F, Thirion B, Varoquaux G (2020) NeuroQuery, comprehensive meta-analysis of human brain mapping. Elife 9 Available at: http://dx.doi.org/10.7554/eLife.53385.

Dosenbach NUF, Koller JM, Earl EA, Miranda-Dominguez O, Klein RL, Van AN, Snyder AZ, Nagel BJ, Nigg JT, Nguyen AL, Wesevich V, Greene DJ, Fair DA (2017) Real-time motion analytics during brain MRI improve data quality and reduce costs. Neuroimage 161:80–93.

Duncan KJ, Pattamadilok C, Knierim I, Devlin JT (2009) Consistency and variability in functional localisers. Neuroimage 46:1018–1026.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25:1325–1335.

Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, Sison ML, Moffitt TE, Caspi A, Hariri AR (2020) What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. Psychol Sci 31:792–806.

Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R (1996) Movement-related effects in fMRI time-series. Magn Reson Med 35:346–355.

Fynes-Clinton S, Marstaller L, Burianová H (2019) Differentiation of functional networks during long-term memory retrieval in children and adolescents. Neuroimage 191:93–103.

Gorgolewski KJ, Storkey A, Bastin ME, Whittle IR, Wardlaw JM, Pernet CR (2013a) A test-retest functional MRI dataset for motor, language and spatial attention functions. Available at: http://gigadb.org/dataset/100051.

Gorgolewski KJ, Storkey AJ, Bastin ME, Whittle I, Pernet CR (2013b) Single subject fMRI test–retest reliability metrics and confounding factors. Neuroimage 69:231–243.

Gratton C, Dworetsky A, Coalson RS, Adeyemo B, Laumann TO, Wig GS, Kong TS, Gratton G, Fabiani M, Barch DM, Tranel D, Dominguez OM-, Fair DA, Dosenbach NUF, Snyder AZ, Perlmutter JS, Petersen SE, Campbell MC (2020a) Removal of high frequency contamination from motion estimates in single-band fMRI saves data without biasing functional connectivity. Neuroimage:116866.

Gratton C, Kraus BT, Greene DJ, Gordon EM, Laumann TO, Nelson SM, Dosenbach NUF, Petersen SE (2020b) Defining Individual-Specific Functional Neuroanatomy for Precision Psychiatry. Biol Psychiatry 88:28–39.

Greene DJ, Koller JM, Hampton JM, Wesevich V, Van AN, Nguyen AL, Hoyt CR, McIntyre L, Earl EA, Klein RL, Shimony JS, Petersen SE, Schlaggar BL, Fair DA, Dosenbach NUF (2018) Behavioral interventions for reducing head motion during MRI scans in children. Neuroimage 171:234–245.

Griffanti L, Douaud G, Bijsterbosch J, Evangelisti S, Alfaro-Almagro F, Glasser MF, Duff EP, Fitzgibbon S, Westphal R, Carone D, Beckmann CF, Smith SM (2017) Hand classification of fMRI ICA noise components. Neuroimage 154:188–205.

Jenkinson M, Bannister PR, Brady JM, Smith SM (2002) Improved optimisation for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM (2012) FSL. Neuroimage 62:782–790.

Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, Davidson RJ, Oakes TR (2006) Motion correction and the use of motion covariates in multiple-subject fMRI analysis.

Hum Brain Mapp 27:779–788.

Kelly AMC, Uddin LQ, Biswal BB, Castellanos FX, Milham MP (2008) Competition between functional brain networks mediates behavioral variability. Neuroimage 39:527–537.

Krause F, Benjamins C, Eck J, Lührs M, van Hoof R, Goebel R (2019) Active head motion reduction in magnetic resonance imaging using tactile feedback. Hum Brain Mapp 40:4026–4037.

Laumann TO, Gordon EM, Adeyemo B, Snyder AZ, Joo SJ, Chen M-Y, Gilmore AW, McDermott KB, Nelson SM, Dosenbach NUF, Schlaggar BL, Mumford JA, Poldrack RA, Petersen SE (2015) Functional System and Areal Organization of a Highly Sampled Individual Human Brain. Neuron 87:657–670.

Lemieux L, Salek-Haddadi A, Lund TE, Laufs H, Carmichael D (2007) Modelling large motion events in fMRI studies of patients with epilepsy. Magn Reson Imaging 25:894–901.

Lewis-Peacock JA, Postle BR (2008) Temporary activation of long-term memory supports working memory. J Neurosci 28:8765–8771.

Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, Hardcastle N, Wexler J, Esteban O, Goncalves M, Jwa A, Poldrack RA (2021) OpenNeuro: An open resource for sharing of neuroimaging data. bioRxiv:2021.06.28.450168 Available at: https://www.biorxiv.org/content/10.1101/2021.06.28.450168v1.full.pdf+html [Accessed July 5, 2021].

Oakes TR, Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, Davidson RJ (2005) Comparison of fMRI motion correction software tools. Neuroimage 28:529–543.

O'Hearn K, Velanova K, Lynn A, Wright C, Hallquist M, Minshew N, Luna B (2016) Abnormalities in brain systems supporting individuation and enumeration in autism. Autism Res 9:82–96.

Patel AX, Kundu P, Rubinov M, Jones PS, Vértes PE, Ersche KD, Suckling J, Bullmore ET (2014) A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. Neuroimage 95:287–304.

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafo MR, Nichols TE, Poline JB, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18:115–126.

Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, Cumba C, Koyejo O, Milham MP (2013) Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform 7:12.

Power JD (2017) A simple but useful way to assess fMRI scan qualities. Neuroimage 154:150–158.

Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012) Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59:2142–2154.

Power JD, Schlaggar BL, Petersen SE (2015) Recent progress and outstanding issues in motion correction in resting state fMRI. Neuroimage 105:536–551.

Power JD, Silver BM, Silverman MR, Ajodan EL, Bos DJ, Jones RM (2019) Customized head molds reduce motion during resting state fMRI scans. Neuroimage 189:141–149.

Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF (2015) ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage 112:267–277.

Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R (2018) Development of the social brain from age three to twelve years. Nat Commun 9:1027.

Rogers CS, Jones MS, McConkey S, Spehar B, Van Engen KJ, Sommers MS, Peelle JE (2020) Age-related differences in auditory cortex activity during spoken word recognition. Neurobiology of Language 1:452–473.

Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM (2014) Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90:449–468.

Satterthwaite TD, Ciric R, Roalf DR, Davatzikos C, Bassett DS, Wolf DH (2019) Motion artifact in studies of functional connectivity: Characteristics and mitigation strategies. Hum Brain Mapp 40:2033–2051.

Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL, Petersen SE (2014) Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. Hum Brain Mapp 35:1981–1996.

Smith SM, Beckmann CF, Ramnani N, Woolrich MW, Bannister PR, Jenkinson M, Matthews PM, McGonigle DJ (2005) Variability in fMRI: a re-examination of inter-session differences. Hum Brain Mapp 24:248–257.

Smyser CD, Snyder AZ, Neil JJ (2011) Functional connectivity MRI in infants: exploration of the functional organization of the developing brain. Neuroimage 56:1437–1452.

Van Dijk KRA, Sabuncu MR, Buckner RL (2012) The influence of head motion on intrinsic functional connectivity MRI. Neuroimage 59:431–438.