An Open Publishing Platform of the Organization for Human Brain Mapping
APERTURE NEURO

ORIGINAL RESEARCH ARTICLE

A Multi-Dataset Evaluation of Frame Censoring for Motion Correction in 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*

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 publicly available datasets representing 11 distinct tasks in child, adolescent, and adult participants. Performance was quantified using maximum *t*-values in group analyses, and region of interest–based mean activation and split-half reliability in single subjects. We compared frame censoring across several thresholds to the use of 6 and 24 canonical motion regressors, wavelet despiking, robust weighted least squares, and untrained ICA-based denoising, for a total of 240 separate analyses. 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 the choice of a motion mitigation strategy depends on both the dataset and the outcome metric of interest.

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

Correspondence: 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, Center for Cognitive and Brain Health, Northeastern University, 360 Huntington, Boston, MA 02115, email: j.peelle@northeastern.edu

Date Received: October 14, 2021

Date Accepted: May 4, 2022

DOI: 10.52294/ApertureNeuro.2022.2.NXOR2026

INTRODUCTION

High-quality neuroimaging analysis depends in part 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, where it can lead to both false positives or a lower signal-to-noise ratio that can make it harder to detect a true activation of interest. As such, motion in task-based fMRI potentially introduces a combination of both Type I and Type II errors.

Rigid body realignment—a mainstay of fMRI analysis for decades—goes some way toward 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 six realignment parameters (translation and rotation around the *X*, *Y*, and *Z* axes, reflecting estimated participant motion) as nuisance regressors in first-level models.

Beyond motion parameter inclusion, 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 that 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. There are different ways to quantify "bad" scans, and choosing both an appropriate metric and associated threshold can also be challenging. Thus, additional information over what threshold should be used for identifying bad frames—and relatedly, how much data are lost versus retained—is necessary to make informed decisions.

Although several published studies compare differing correction strategies (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 participant populations. Although our primary interest was frame censoring, we considered seven different motion correction approaches:

1. six canonical head motion (i.e., "realignment parameter") 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 (uICA)
6. frame censoring based on frame displacement (FD)
7. frame censoring based on variance differentiation (DVARS)

This list is not exhaustive but is 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 four complementary outcome metrics: (1) the maximum group *t*-statistic both across the whole brain and in a region of interest (ROI) relevant to the task; (2) the average parameter estimates from within the same ROI; (3) the degree of test–retest consistency exhibited by subject-level parametric maps; and (4) the spatial overlap of thresholded group-level statistical maps. These metrics are simple to define yet functionally meaningful and can be applied to data from almost any fMRI study. In our view, Dice quantifies replicability, the mean ROI value quantifies effect size (signal), and maximum-*t* quantifies signal to noise (effect size penalized by variance).

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 are provided in **Supplemental Table 2**). Additional information, including task details, modeling/contrast descriptions compiled from

Table 1. Summary of datasets analyzed

Dataset	Reference	Task and design	# subs	Age range	FD (median ± SD)	Frames per subject
ds000102	Kelly et al. (2008)	Flanker (E)	22	22–50	0.11 ± 0.12	284
ds000107	Duncan et al. (2009)	1-back (B)	43	19–38	0.08 ± 0.14	323
ds000114	Gorgolewski et al. (2013b)	Motor (B)	10	50–58	0.14 ± 0.16	360
		Covert verb (B)	10	50–58	0.11 ± 0.11	338
		Overt word (B)	10	50–58	0.13 ± 0.12	144
		Line bisection (B)	9	50–58	0.13 ± 0.18	468
ds000228	Richardson et al. (2018)	Film viewing (E)	122	3.5–12	0.21 ± 0.93	164
			33	18–39	0.18 ± 0.27	164
ds001497	Lewis-Peacock and Postle (2008)	Face perception (E)	10	19–32	0.11 ± 0.12	1146
ds001534	Courtney et al. (2018)	Food images (E)	42	18–22	0.10 ± 0.16	552
ds001748	Fynes-Clinton et al. (2019)	Memory retrieval (E)	21	10–12	0.16 ± 0.36	438
			20	14–16	0.12 ± 0.17	438
			21	20–35	0.08 ± 0.17	438
ds002382	Rogers et al. (2020)	Word recognition (E)	29	19–30	0.14 ± 0.35	710
			32	65–81	0.30 ± 0.34	710

B = blocked design; E = event-related design.

publication(s) associated with a given study, and any data irregularities encountered during analysis, is provided in the **Supplemental Materials**.

Analysis

All scripts used in the study are available at https://osf.io/n5v3w/. 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 the 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.

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 to 2 mm isotropic voxels and smoothing of the functional images using an 8-mm FWHM Gaussian kernel.

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) untrained ICA denoising, (6) frame censoring based on framewise displacement (FD), or (7) differential variance (DVARS) thresholding (FD/DVARS thresholding is described later).

Statistical modeling was performed in SPM for 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 untrained ICA denoising. As recommended by developers of the rWLS toolbox, unsmoothed data were used for variance estimation and contrast maps were smoothed after modeling. For wavelet despiking, functional images were

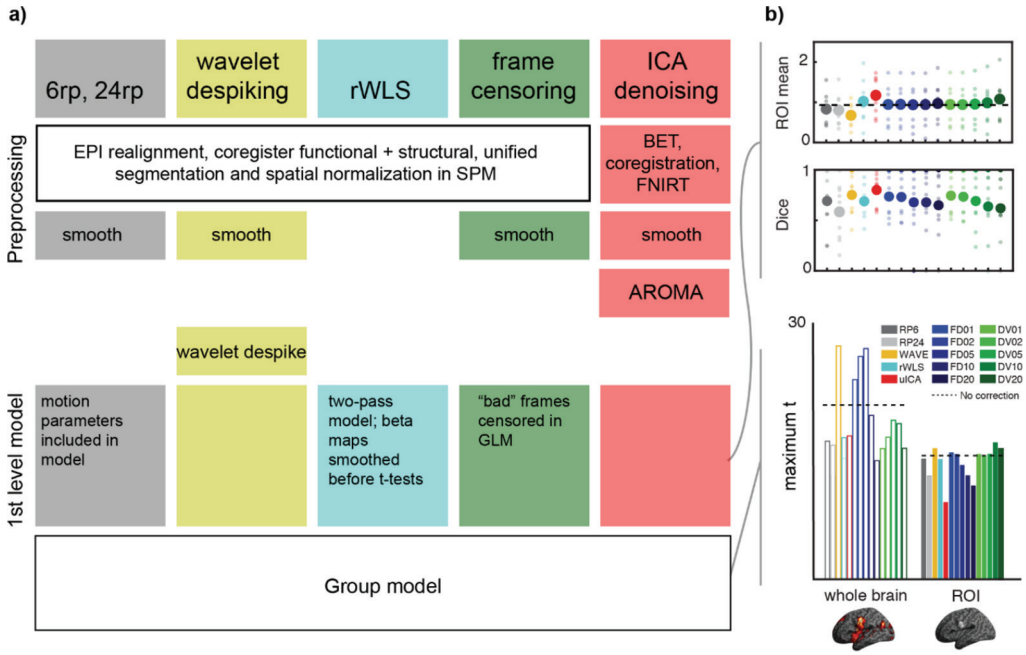

Fig. 1. Schematic of processing pipeline and outcome measures. **(a)** Summary of preprocessing and model steps in common and differing across motion correction strategies. **(b)** Following statistical modeling, outcomes are summarized in mean parameter estimates and Dice overlap of thresholded single-subject maps (top) and maximum t-value from the group analysis (bottom). Dashed lines represent values obtained without motion correction.

rescaled to a whole-brain median of 1000 across all frames before processing. The default toolbox settings (wavelet: d4, threshold: 10, boundary: reflection, chain search: moderate, scale number: liberal) were used. Finally, untrained 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. This ICA implementation is not based on training data, and so we refer to it as "untrained" ICA to distinguish it from other ICA-based denoising approaches.

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, multisession data were treated as a test–retest paradigm (the first session statistical map was compared to the second session in studies having fewer than three sessions; even-numbered versus odd-numbered sessions were compared otherwise). Replicability was quantified as the Dice coefficient of thresholded first-level *t*-maps ($p < 0.001$, 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 (**Supplemental Figure S1**). 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,

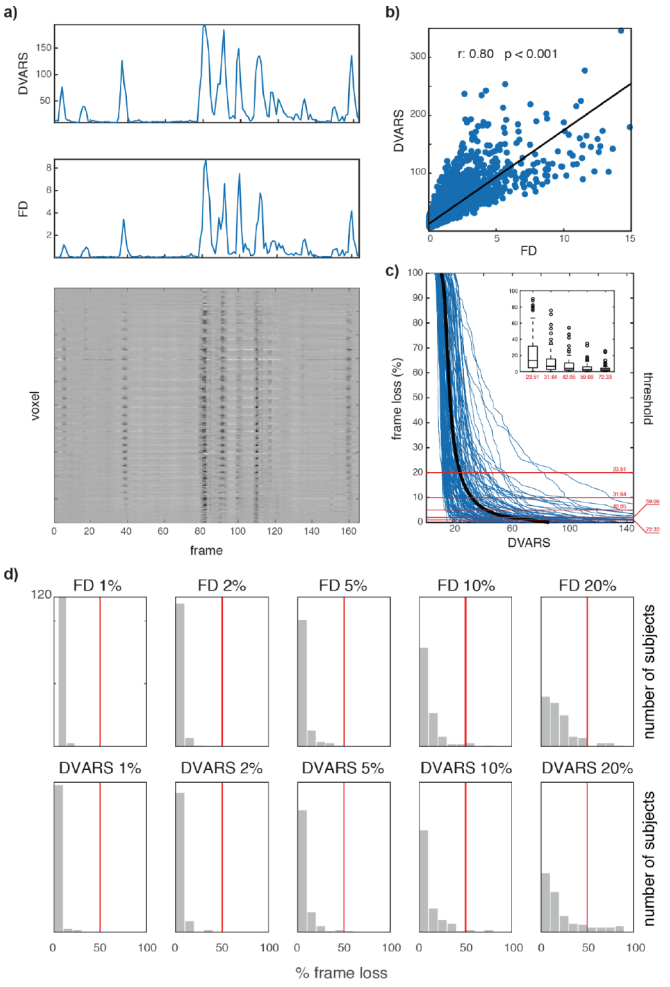

Fig. 2. Calculation of censoring thresholds in an example dataset (ds000228, children). **(a)** Representative grayplot (*bottom;* Power 2017) showing 500 randomly selected gray matter voxels. DVARS and FD for this session are plotted above. Spikes in the metrics identify frames contaminated by artifacts. **(b)** DVARS and FD are correlated but exhibit differing amplitudes and time courses. As such, the use of both measures was explored. **(c)** Metric values (here shown for DVARS) used for censoring were determined by plotting frameloss for each subject as a function of threshold (thin blue traces). Interpolation of the mean response (thick black trace) gives metric values corresponding to a data loss of 1%, 2%, 5%, 10%, or 20%. Box plot (*inset*) summarizes the results across all subjects at each threshold (box: 25–75% percentiles; crosses: >3 SD outliers). **(d)** Histograms of data loss at each FD or DVARS target value across all subjects in this dataset. Red line indicates 50% frameloss.

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). Second, 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 single threshold would not be suitable for all datasets. We, therefore, employed the frame-percent thresholding strategy to obtain a reasonable range of results in all studies examined. To be clear, we do not propose a fixed-percent data loss approach as a "production" strategy. Indeed, a conventional denoising approach would be to select an absolute threshold of acceptable motion based on experience (or preliminary data) from the target subject pool, task, scanner, and so on. However, given the variety of datasets examined here, we had no *a priori* guide as to what threshold values to use. Any fixed selection might censor no frames in some studies and too many in others. Therefore, we employed the frame-percent thresholding strategy to obtain an informative range of results in all datasets. The threshold values that resulted from percent data loss targeting in these datasets are shown in **Supplemental Figure S2** and listed in **Supplemental Table 3**. The amount of data censored in each participant in a single study is shown in **Figure 2d**, and for all studies in **Supplemental Figure S3**.

To implement frame censoring, first-level modeling was repeated for each threshold with a separate 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. Although some prior studies of motion correction have censored one or more frames before or following threshold violations (e.g., "augmentation" of Siegel et al. 2014), we did not pursue such variations to avoid further expanding what was already a rather large parameter space.

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 probability map in the SPM Anatomy Toolbox (Eickhoff et al. 2005) for a task-relevant brain structure or anatomical region (e.g., "V2"). Additional details on the ROI definition used in each analysis are provided in the **Supplemental Materials**.

RESULTS

Performance of the motion correction strategies organized by dataset is shown in **Figure 3**. Each panel includes a 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.

These results show there is a 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 tends 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-constrained

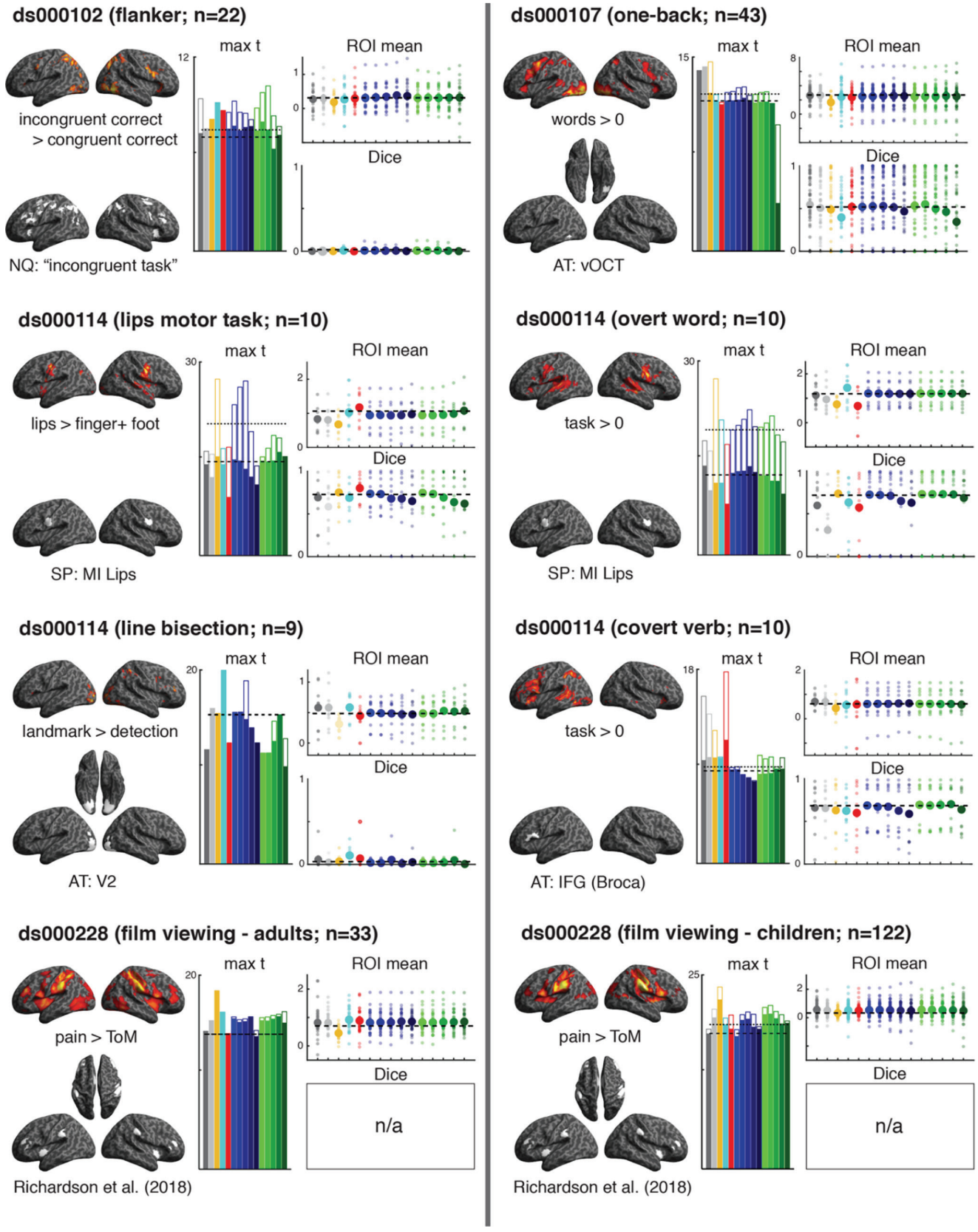

Fig. 3. Summary of motion correction algorithm performance for all datasets examined in the study. Each panel includes a representative thresholded group *t*-map at left (p = 0.001, uncorrected) for the given contrast with the ROI used for evaluation plotted below (AT = Anatomy Toolbox probability map; NQ = NeuroQuery search term; SP = 5 mm sphere centered at the described landmark. The ROI used in the analysis of ds000228 is defined in Table 2 of Richardson et al. [2018]). At the center, ROI-restricted maximum *t*-values are superimposed on whole-brain results for each motion correction 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." Horizontal reference lines indicate the value obtained when no motion correction was used (dashed: ROI-constrained; dotted: whole brain).

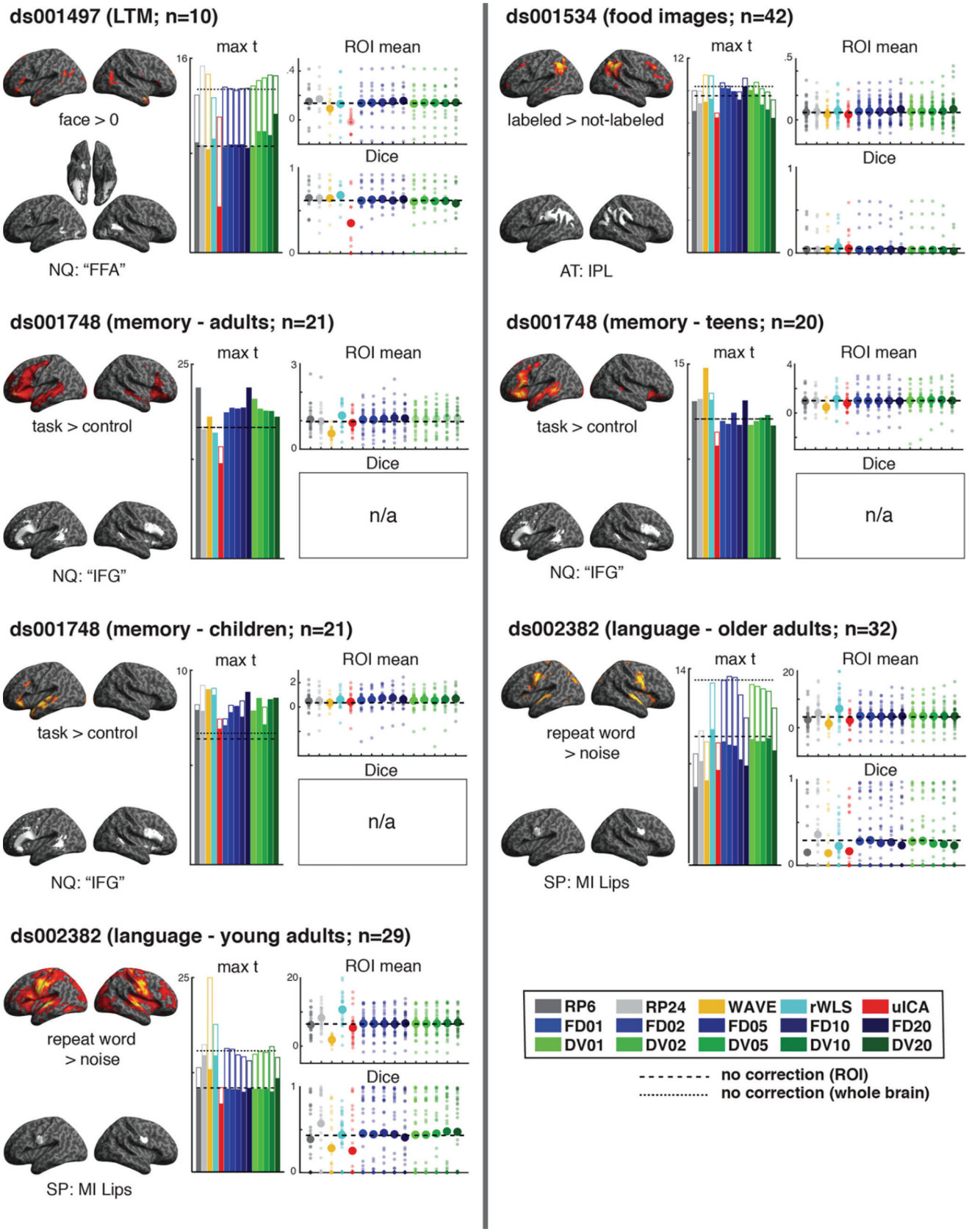

Fig. 3. (*continued*)

maximum *t*-value, shown in filled bars, are superimposed on the whole-brain results, shown in open bars in **Figure 3**). Conversely, untrained ICA gives consistently poorer results although it offers the best maximum *t*-value in the ds000114 covert verb task. Performance of FD and DVARS frame censoring was highly variable, with the application of increasingly stringent thresholds improving performance in some datasets while decreasing it in others. A somewhat consistent result 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.

The mean effect size shown in these results was largely insensitive to the selected motion correction approach. The two exceptions are wavelet despiking and untrained ICA, which produce consistently smaller values than the other approaches. This may reflect suboptimal parameter selection in these algorithms (see **Discussion**). Robust weighted least squares offers 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 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, all of which unfortunately provided an uninformative test–retest quantification, some trends can be identified. There is a decrease in both the FD and DVARS frame censoring results, especially at 20% thresholding. In general, all differences were minor, save for untrained ICA which performs notably better in the ds000114 motor task and notably worse in ds001487. The reason why three datasets exhibit poor performance in a test–retest paradigm is unclear. Although ds000114 had a relatively small subject pool (n = 10), both ds000102 and ds001534 used a larger sample size (n = 22 and n = 42, respectively). Whatever the cause, it appears to be unrelated to the choice of motion correction, as in these exceptions all strategies performed equally well (or equally poorly, as it were).

A summary of univariate results is shown in **Figure 4a**, in which mean values of all four performance metrics are plotted. Several of the trends noted in the individual datasets remain apparent. For example, wavelet despiking gave the largest whole-brain maximum *t*-value, whereas robust weighted least squares resulted in the best ROI-constrained performance. Light-to-moderate frame censoring resulted in improvement which then declined as more aggressive thresholding was applied. Robust weighted least squares produced the largest average effect size. Wavelet despiking and untrained ICA

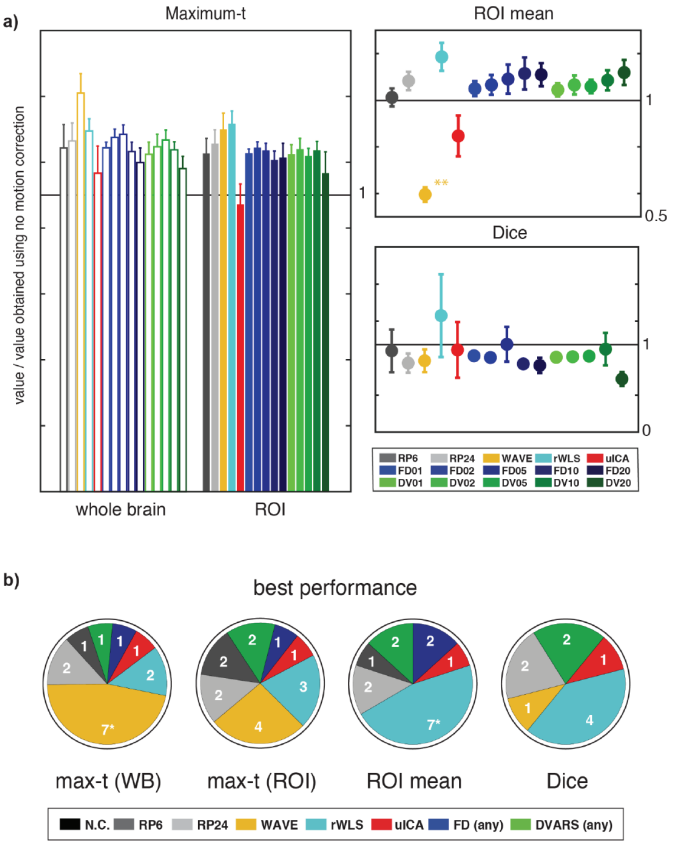

Fig. 4. Summary statistics. **(a)** Whole-brain and ROI-restricted maximum *t*-values (*left*) mean effect size (*upper right*), and test–retest Dice coefficient (*lower right*) averaged across all datasets. Values were normalized to the no-correction values in the same dataset before pooling. **(b)** Performance summarized as a count of the datasets a given approach gave best results measured by the four performance metrics (*$p < 0.05$).

produce poor results as measured by this metric. Test–retest performance is generally poorer for most motion correction strategies than that obtained using no motion correction, although rWLS exhibits good performance as measured by this metric.

An omnibus ANOVA identified a significant difference in the maximum-*t* data; however, Scheffe post hoc testing found no significant pairwise differences ($p > 0.05$). Both omnibus and post hoc testing of the mean ROI effect size show wavelet despiking differed significantly from all other approaches ($p < 0.001$). No significant differences were found in the test–retest Dice data.

A count summary of best algorithm performance is shown in **Figure 4b**, in which the best performing motion correction approach for each metric was identified in each of the 15 datasets, and the resulting proportions plotted as pie charts. The general trends evident in the averaged results are also apparent in these data although some additional features emerge. Robust weighted least squares offered the best performance on many datasets. Wavelet despiking gave the best maximum *t*-value in approximately half (whole-brain) or one quarter (ROI-constrained) of the studies. Untrained ICA gave the best results across all four metrics in at least one dataset.

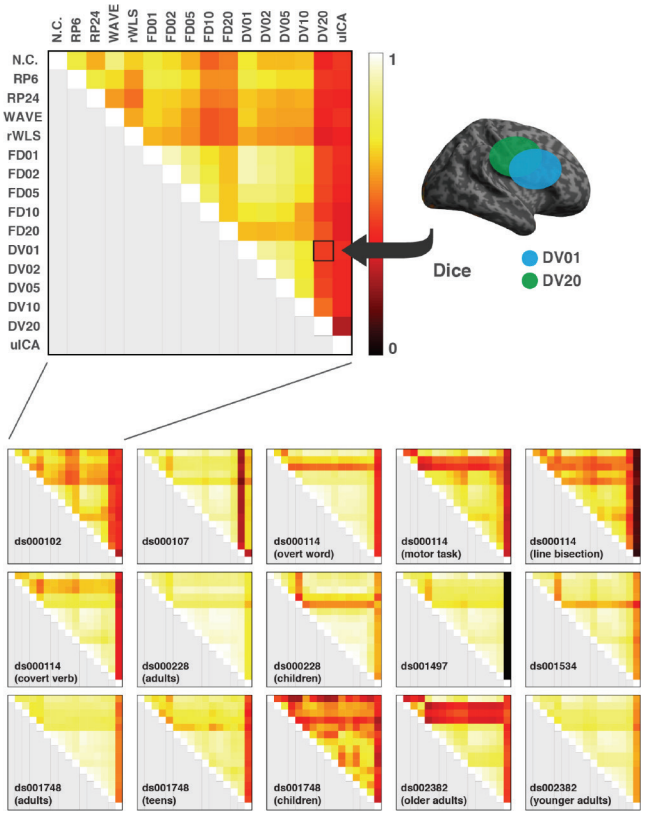

Fig. 5. Quantification of overlap in group-level maps. Dice coefficients computed from group-level thresholded statistical maps obtained using each pair of motion correction strategies were assembled into a 15 × 15 Dice matrix. Overlap shown here for DV01 and DV20 is illustrative.

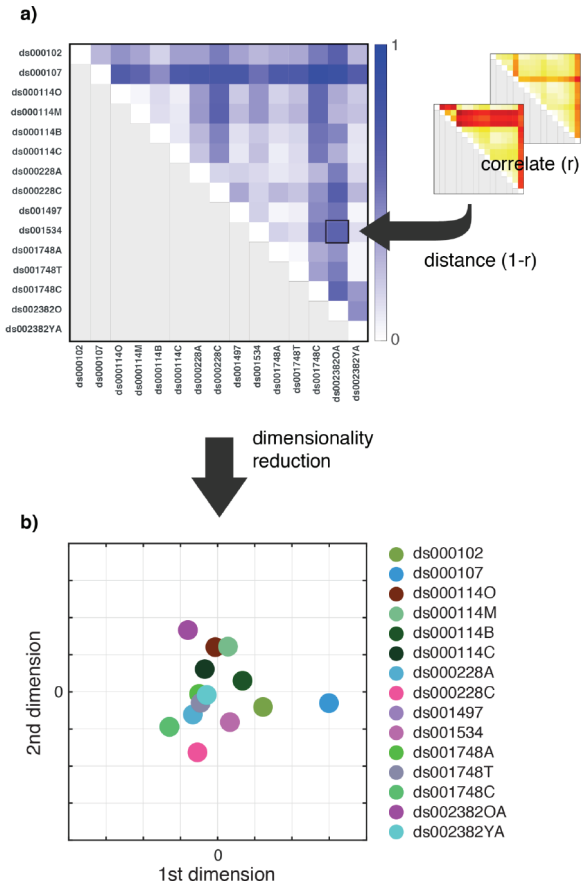

Fig. 6. Multivariate analysis of group-level overlap. (a) Representational dissimilarity matrix (RDM) illustrating the distance between motion correction patterns for each of the 15 datasets shown in Figure 5. **(b)** Using multidimensional scaling (MDS), we visualized the relative distances between datasets in a reduced number of dimensions. Plotting the first two dimensions partially segregates ds000107 (cyan dot) but does not suggest other organizations of the datasets. Plots of other low-dimensional projections were qualitatively similar (see also **Supplemental Figure S4**).

Frame censoring performed similarly using either FD or DVARS. Finally, the performance of the RP6 and RP24 approaches are middling, producing the best maximum *t*-value on only one or two datasets and, with one exception, never producing the best ROI mean or test–retest results. However, of these results, only the maximum-*t* performance of wavelet despiking and rWLS ROI mean effect size were statistically significant ($p < 0.05$).

Given the substantial variability in motion correction results across datasets, we next explored whether there may have been systematic differences between datasets that affected motion correction performance. We first calculated the pairwise similarity of thresholded (voxelwise $p < 0.001$) group maps from each dataset using Dice overlap **(Figure 5)**. A consistent finding was a generally lower overlap between untrained ICA and the other motion correction approaches. Additionally, RP6, RP24, and rWLS tended to overlap less with other motion correction approaches and more with one another, although exceptions can be noted. Results for most datasets are generally mixed, although ds000228 (adults), ds001748 (adults), and ds002382 (young adults) exhibit high overlap for all motion correction approaches (with the exception of untrained ICA).

Having generated Dice overlap maps for each dataset, we then explored the higher-order relationship between datasets using representational similarity analysis (Kriegeskorte et al. 2008). We first calculated the distance between each Dice matrix using Pearson correlation, creating a representational dissimilarity matrix (RDM) based on these distances (**Figure 6a**). We then used multidimensional scaling (MDS) to visualize the relationship between datasets. A plot of the data in the first two eigen dimensions is shown in **Figure 6b**. Dataset ds000107 appears at the right edge of the space, as might be predicted by a visual review of the RDM. However, the other datasets present no distinct pattern. A plot of the data using the first three dimensions similarly exhibited no distinct features, as did an examination of all 2D projections using the first five eigen-dimensions (see **Supplemental Material Figure S4**).

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, 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. The 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. 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 the 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, the 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. 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 the 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. 2013a; Elliott et al. 2020), at least under conventional scanning protocols (Gratton et al. 2020b).

Comparing motion correction approaches

No single motion correction 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 we could identify.

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—performed 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 poorer 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 the best performance in approximately half of the datasets across all performance metrics (**Figure 4b**). 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 differentially weight time points, reducing the influence of motion on parameter estimates. However, we found that other motion correction strategies offered 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 not optimally) remove all of the artifacts.

In contrast to the 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 toolbox implementation. It should also be noted the wavelet despiking 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 parameter estimate). Thus, numeric comparison to other approaches (that do not include rescaling) is problematic. It also may be possible to improve performance—including obtaining effect sizes concomitant with other motion correction approaches—by tuning the algorithm.

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, it should be emphasized that we implemented ICA denoising using FSL's ICA-AROMA, specifically 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 presently because we sought to evaluate tools for motion correction that could be implemented within a completely automated pipeline. The potential of ICA, in general, 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

Frame censoring has appeared in several 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 correction was used. This suggested that using a discrete rather than a continuous nuisance regressor may better preserve task-related 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 data quality. Because 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 or by the effective loss of task-related data. 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).

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 somewhat mixed support for such a position. We investigated applying a (fixed) FD threshold of 0.9 to these data (**Supplemental Figure S2**), as used by Siegel and colleagues (2014) in their exploration of frame censoring and as well as other 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 are optimal 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 that there is no single threshold value that will optimize frame censoring for all datasets and outcome measures. Although for individual investigators it may indeed make more sense to develop censoring criteria based on the range of FD or DVARS values present in their specific data, we also suggest that considering the amount of data lost at a chosen threshold is a useful metric to take into consideration.

Effects of FD-based versus DVARS-based thresholding

A consistent finding in the present study was that different frame censoring outcomes are obtained depending on whether FD or DVARS is used for thresholding. This effect is most striking in the maximum *t*-values observed in the individual studies (**Figure 3**). Systematically varying the FD and DVARS threshold values resulted in dissimilar or even contrary effects, with improvements observed in one metric often contrasting with worsening performance in the other. Although perhaps unexpected at first glance, this result reflects the nature of the two parameters and how censored frames are identified.

While FD is a direct quantification of estimated head motion, DVARS is potentially affected by any process that changes image intensity between frames. This includes not only head motion but also both neural and non-neural influences such as arousal (Gu et al. 2020), respiration (Power et al. 2018), and cerebrospinal fluid flow (Fultz et al. 2019). As a result, even though FD and DVARS are strongly correlated, they are not identical, and this disparity is responsible for the observed differences in FD and DVARS performance. Even if the

number of censored frames is equivalent (cf. **Figure 2d**), a different collection of frames is targeted by each parameter at a given threshold. The relationship between FD- and DVARS-based thresholding can be conveniently demonstrated by considering the scatterplot of FD versus DVARS in **Figure 2b**. FD thresholding can be viewed in this plot as a vertical line moving right to left as the threshold is made more stringent. On the other hand, DVARS thresholding corresponds to a horizontal line moving top to bottom. Although there is a general overlap in the frames that violate both thresholds, the collections are not identical. Because the relation between the two parameters differs in each dataset (see **Supplemental Figure S3**), different trends in FD- and DVARS-based thresholded performance emerge.

Patterns of results across datasets

The similarity analysis of group-level maps (**Figure 5**) exhibits several notable features. First, untrained ICA had relatively low overlap with other motion correction strategies in most (but not all) datasets. Despite the frequently lower Dice scores, we did not see results for untrained ICA that were substantially mismatched with the other results. A review of the data reveals that the performance of untrained ICA seemed to result from less-extensive activation compared to group-level maps obtained using the other motion correction approaches. Stated differently, the untrained ICA activation maps were not "incorrect"; they were simply more focal (and thus overlapped less with other approaches).

Second, RP6 and RP24 produced a lower Dice overlap in many datasets. As these techniques are based on the use of continuous regressors, they represent an algorithmically distinct approach compared to temporally compact (wavelet) or discrete regressors (frame censoring). This effect can also be seen in the results of robust weighted least squares, which in some datasets (e.g., ds001497 and ds001534) produce the only notable Dice difference. As such, a tempting takeaway is that the motion correction strategies based on continuous regressors form a performance family. However, when all performance metrics are considered collectively, the distinction between approaches becomes less clear.

Finally, some of the overlap performance appears to be related to data quality. For example, ds001748 and ds002382 explored identical tasks across multiple samples of approximately equal size. Both datasets included a high-motion group (the children group in ds0001748 and the older adults in ds002382—see **Table 1**), and it is these Dice matrices that exhibit the greatest variability within the group. Conversely, the Dice matrices for the ds001748 adult and teen subject pools and the young adults in ds002382 are relatively uniform. This suggests that the choice of a motion correction strategy may be less important when working with a subject pool exhibiting only minor motion, at least when considering the spatial distribution of group-level activation.

These qualitative differences suggest Dice overlap might offer a means of categorizing the datasets and in so doing might provide a guideline for the selection of a motion correction strategy. A five-group categorization of the datasets can be proposed based simply on their appearance in Figure 5: (1) ds000102 and ds000114 (line bisection), (2) ds000114 (motor) and ds002382 (older adults), (3) ds000228 (adults), ds001748 (adults), and ds002382 (younger adults), (4) ds001748 (children), and (5) all remaining datasets. Yet, the quantitative results of our RDM-informed multidimensional scaling (**Figure 6** and **Supplementary Figure S4**) do not support this organization. Our goal was to identify common features of datasets using the overall pattern of motion correction results, which we operationalized using Dice overlap. However, this was not the case: MDS was unable to reduce the dimensionality of these data in a way that supplied meaningful information, and studies grouped together even using the informal visual organization described earlier differ in subject pools, task type, and other characteristics. Like the univariate metrics considered here, our multivariate analysis failed to clearly identify characteristics that might be used to identify an optimal motion correction strategy. It could be that a similar approach, but with hundreds of data sets, would be able to identify systematic differences in how different motion correction strategies worked on different types of data, which may be a promising direction for future work.

Other considerations

We have focused on retrospective correction—that is, strategies for dealing with motion in the existing data. A complementary approach would be to reduce head motion during acquisition. Protocols have been developed to do so, 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 motion correction) will follow.

A possibility that we did not explore is combining strategies, as is commonly done in resting-state fMRI (e.g., frame censoring of outliers followed by including motion regressors from rigid-body realignment). However, this expands an already unwieldy parameter space of possible analyses (Carp 2012; Poldrack et al. 2017; Botvinik-Nezer et al. 2020). The use of simulated data, where "ground truth" can be known, may also prove beneficial in understanding how motion correction strategy can affect the validity of our inferences.

Conclusions

The present results do not identify unequivocal guidelines for selecting a motion correction strategy. Given the variability observed across datasets, analyzed using identical processing pipelines, exploring multiple strategies in a given dataset may be the best way of reducing motion artifacts. Although it may be possible to revisit this issue in future work, our present results suggest that—frustratingly—no single motion correction strategy will give optimal results in every instance, 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, R01 DC019507, and T32 EB014855 from the US National Institutes of Health. OpenNeuro is supported by NSF Grant OCI-1131441.

REFERENCES

1. Ardekani BA, Bachman AH, Helpern JA. A quantitative comparison of motion detection algorithms in fMRI. Magn Reson Imaging. 2001 Sep;19(7):959–63.
2. Ashburner J, Friston KJ. Rigid Body Registration. In: Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Price CJ, Zeki S, et al., editors. Human Brain Function. 2nd ed. New York: Elsevier; 2004. p. 635–53.
3. Ashburner J, Friston KJ. Unified segmentation. Neuroimage. 2005 Jul 1;26(3):839–51.
4. Bakkour A, Lewis-Peacock JA, Poldrack RA, Schonberg T. Neural mechanisms of cue-approach training. Neuroimage. 2017 May 1;151:92–104.
5. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci. 2010 Mar;1191:133–55.
6. Bianciardi M, Fukunaga M, van Gelderen P, Horovitz SG, de Zwart JA, Shmueli K, et al. Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. Magn Reson Imaging. 2009 Oct;27(8):1019–29.
7. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. Nature. 2020 Jun;582(7810):84–8.
8. Carp J. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front Neurosci. 2012 Oct 11;6:149.
9. Courtney AL, PeConga EK, Wagner DD, Rapuano KM. Calorie information and dieting status modulate reward and control activation during the evaluation of food images. PLoS One. 2018 Nov 2;13(11):e0204744.
10. Cusack R, Vicente-Grabovetsky A, Mitchell DJ, Wild CJ, Auer T, Linke AC, et al. Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. Front Neuroinform. 2015;8:90.
11. Davis T, Goldwater M, Giron J. From concrete examples to abstract relations: the rostrolateral prefrontal cortex integrates novel examples into relational categories. Cereb Cortex. 2017 Apr 1;27(4):2652–70.
12. Diedrichsen J, Shadmehr R. Detecting and adjusting for artifacts in fMRI time series data. Neuroimage. 2005;27:624–34.
13. Dockès J, Poldrack RA, Primet R, Gözükan H, Yarkoni T, Suchanek F, et al. NeuroQuery, comprehensive meta-analysis of human brain mapping. Elife [Internet]. 2020 Mar 4;9. Available from: http://doi.org/10.7554/eLife.53385
14. Dosenbach NUF, Koller JM, Earl EA, Miranda-Dominguez O, Klein RL, Van AN, et al. Real-time motion analytics during brain MRI improve data quality and reduce costs. Neuroimage. 2017;161:80–93.
15. Duncan KJ, Pattamadilok C, Knierim I, Devlin JT. Consistency and variability in functional localisers. Neuroimage. 2009;46(4):1018–26.
16. Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, et al. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage. 2005 May 1;25(4):1325–35.
17. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, et al. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol Sci. 2020 Jul;31(7):792–806.
18. Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R. Movement-related effects in fMRI time-series. Magn Reson Med. 1996 Mar;35(3):346–55.
19. Fultz NE, Bonmassar G, Setsompop K, Stickgold RA, Rosen BR, Polimeni JR, et al. Coupled electrophysiological, hemodynamic, and cerebrospinal fluid oscillations in human sleep. Science. 2019 Nov 1;366(6465):628–31.
20. Fynes-Clinton S, Marstaller L, Burianová H. Differentiation of functional networks during long-term memory retrieval in children and adolescents. Neuroimage. 2019 May 1;191:93–103.
21. Gorgolewski KJ, Storkey AJ, Bastin ME, Whittle IR, Pernet CR. Single subject fMRI test–retest reliability metrics and confounding factors. Neuroimage. 2013a;69:231–43.
22. Gorgolewski KJ, Storkey AJ, Bastin ME, Whittle IR, Wardlaw JM, Pernet CR. A test-retest functional MRI dataset for motor, language and spatial attention functions [Internet]. GigaScience Database; 2013b. Available from: http://gigadb.org/dataset/100051
23. Gratton C, Dworetsky A, Coalson RS, Adeyemo B, Laumann TO, Wig GS, et al. Removal of high frequency contamination from motion estimates in single-band fMRI saves data without biasing functional connectivity. Neuroimage. 2020a Apr 20;116866.
24. Gratton C, Kraus BT, Greene DJ, Gordon EM, Laumann TO, Nelson SM, et al. Defining individual-specific functional neuroanatomy for precision psychiatry. Biol Psychiatry. 2020b Jul 1;88(1):28–39.
25. Greene DJ, Koller JM, Hampton JM, Wesevich V, Van AN, Nguyen AL, et al. Behavioral interventions for reducing head motion during MRI scans in children. Neuroimage. 2018;171:234–45.
26. Griffanti L, Douaud G, Bijsterbosch J, Evangelisti S, Alfaro-Almagro F, Glasser MF, et al. Hand classification of fMRI ICA noise components. Neuroimage. 2017 Jul 1;154:188–205.
27. Gu Y, Han F, Sainburg LE, Liu X. Transient arousal modulations contribute to resting-state functional connectivity changes associated with head motion parameters. Cereb Cortex. 2020 Sep 3;30(10):5242–56.
28. Jenkinson M, Bannister PR, Brady JM, Smith SM. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. Neuroimage. 2002;17(2):825–41.
29. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. Neuroimage. 2012 Aug 15;62(2):782–90.
30. Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, Davidson RJ, et al. Motion correction and the use of motion covariates in multiple-subject fMRI analysis. Hum Brain Mapp. 2006 Oct;27(10):779–88.
31. Kelly AMC, Uddin LQ, Biswal BB, Castellanos FX, Milham MP. Competition between functional brain networks mediates behavioral variability. Neuroimage. 2008 Jan 1;39(1):527–37.
32. Krause F, Benjamins C, Eck J, Lührs M, van Hoof R, Goebel R. Active head motion reduction in magnetic resonance imaging using tactile feedback. Hum Brain Mapp. 2019 Oct 1;40(14):4026–37.
33. Lemieux L, Salek-Haddadi A, Lund TE, Laufs H, Carmichael D. Modelling large motion events in fMRI studies of patients with epilepsy. Magn Reson Imaging. 2007 Jul;25(6):894–901.
34. Lewis-Peacock JA, Postle BR. Temporary activation of long-term memory supports working memory. J Neurosci. 2008 Aug 27;28(35):8765–71.
35. Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, et al. The OpenNeuro resource for sharing of neuroscience data. eLife. 2021 Oct 18;10:e71774.
36. Oakes TR, Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, et al. Comparison of fMRI motion correction software tools. Neuroimage. 2005 Nov 15;28(3):529–43.
37. O'Hearn K, Velanova K, Lynn A, Wright C, Hallquist M, Minshew N, et al. Abnormalities in brain systems supporting individuation and enumeration in autism. Autism Res. 2016 Jan;9(1):82–96.
38. Patel AX, Kundu P, Rubinov M, Jones PS, Vértes PE, Ersche KD, et al. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. Neuroimage. 2014 Jul 15;95:287–304.
39. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafo MR, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci. 2017 Feb;18(2):115–26.

40. Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, et al. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform. 2013;7:12.
41. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage. 2012;59:2142–54.
42. Power JD, Plitt M, Gotts SJ, Kundu P, Voon V, Bandettini PA, et al. Ridding fMRI data of motion-related influences: removal of signals with distinct spatial and physical bases in multiecho data. Proc Natl Acad Sci U S A. 2018 Feb 27;115(9):E2105–14.
43. Power JD, Schlaggar BL, Petersen SE. Recent progress and outstanding issues in motion correction in resting state fMRI. Neuroimage. 2015 Jan 15;105:536–51.
44. Power JD, Silver BM, Silverman MR, Ajodan EL, Bos DJ, Jones RM. Customized head molds reduce motion during resting state fMRI scans. Neuroimage. 2019 Apr 1;189:141–9.
45. Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage. 2015 May 15;112:267–77.
46. Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. Development of the social brain from age three to twelve years. Nat Commun. 2018 Mar 12;9(1):1027.
47. Rogers CS, Jones MS, McConkey S, Spehar B, Van Engen KJ, Sommers MS, et al. Age-related differences in auditory cortex activity during spoken word recognition. Neurobiol Lang. 2020 Oct;1(4):452–73.
48. Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage. 2014 Apr 15;90:449–68.
49. Satterthwaite TD, Ciric R, Roalf DR, Davatzikos C, Bassett DS, Wolf DH. Motion artifact in studies of functional connectivity: characteristics and mitigation strategies. Hum Brain Mapp. 2019 May;40(7):2033–51.
50. Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL, et al. Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. Hum Brain Mapp. 2014;35:1981–96.
51. Smith SM, Beckmann CF, Ramnani N, Woolrich MW, Bannister PR, Jenkinson M, et al. Variability in fMRI: a re-examination of inter-session differences. Hum Brain Mapp. 2005 Mar;24(3):248–57.
52. Smyser CD, Snyder AZ, Neil JJ. Functional connectivity MRI in infants: exploration of the functional organization of the developing brain. Neuroimage. 2011 Jun 1;56(3):1437–52.
53. Van Dijk KRA, Sabuncu MR, Buckner RL. The influence of head motion on intrinsic functional connectivity MRI. Neuroimage. 2012 Jan 2;59(1):431–8.

Supplemental Methods

Representational similarity analysis of group-level maps. Overlap of the thresholded group-level maps ($p < 0.001$, uncorrected) was quantified using a Dice coefficient computed for all pairs of denoising approaches. The result was a 16 × 16 matrix for each dataset in which the (i,j)-th entry is the Dice coefficient quantifying overlap of the group-level map obtained using the i-th denoising strategy with that of the j-th denoising strategy. The overlap summaries were then used to explore a multivariate characterization of the denoising strategies. To this end, all pairs of the 15 dice overlap matrices were correlated and the coefficients were collected into a 15 × 15 distance matrix (distance = 1 − r). Multidimensional scaling (MDS; Matlab function *cmdscale*) was then applied to the distance matrix to obtain a low-dimensional approximation which was plotted and examined for clustering or other patterns.

Univariate statistical analysis. Maximum t values, mean ROI effect size, and Dice test–retest values for each denoising strategy were pooled across all datasets and analyzed using a repeated-measures analysis of variance (Matlab functions *fitrm* and *rmanova*). Maximum t, effect size, and Dice test–retest score were normalized to the value obtained using no motion correction prior to pooling to account for variability across datasets. Pairwise differences were identified in post hoc testing using a Scheffe test (Matlab function *multcompare*).

An alternate evaluation of algorithm performance was explored by simply counting the number of datasets each approach gave the best result (i.e., the largest value) on a given metric. For the purpose of this analysis, FD and DVARS results were combined across all percent data loss categories, reducing the total number of categories to eight. Counts were then evaluated statistically using a binomial test (Zar 2010). Under the null hypothesis that all denoising strategies perform equally well, exhibiting the best performance on seven or more datasets by any one strategy is significant at the 0.05 level.

Supplemental Results

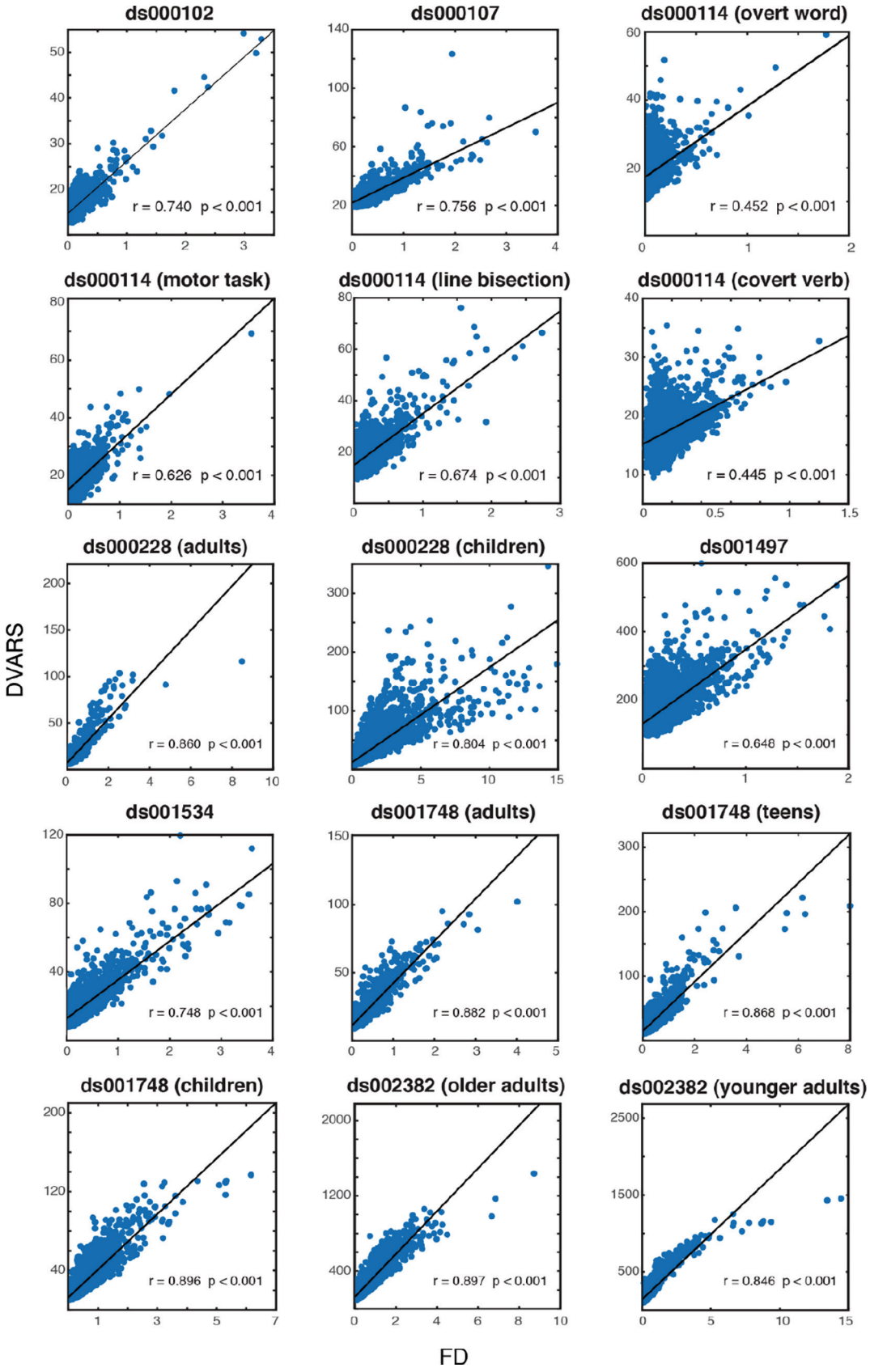

Fig. S1. Correlation of FD and DVARS for all datasets. Each point is the FD (*x*) and corresponding DVARS (*y*) for a single frame, combined across all sessions and subjects for a given dataset. Each plot includes a least-squares line fit to the data.

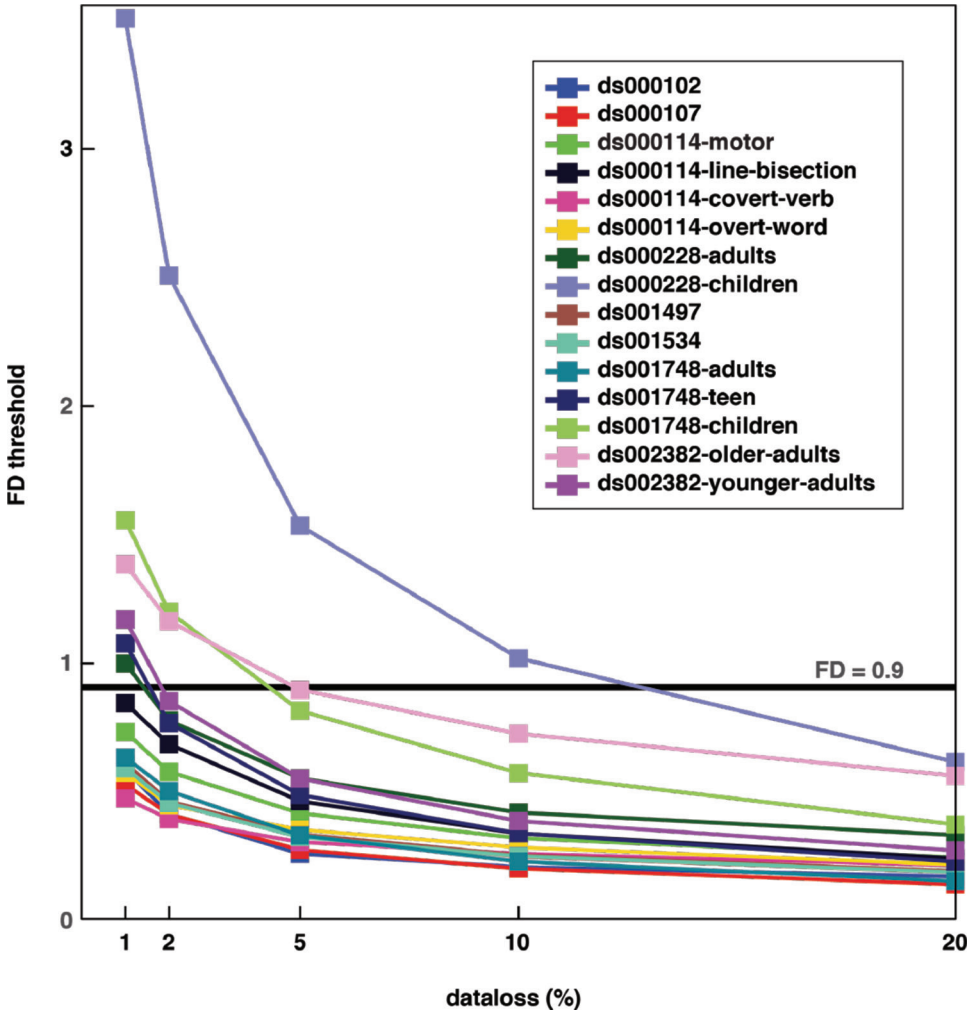

Fig. S2. FD values resulting from targeted percent data loss. The FD threshold value that resulted in 1%, 2%, 5%, 10%, or 20% data loss for a given dataset can be read from the *y*-axis. A horizontal line at FD = 0.9 is included to illustrate the data loss that would have occurred had an (arbitrary) fixed FD threshold of 0.9 been applied (given by the intersection or extrapolated intersection of the horizontal line with the graph of a dataset).

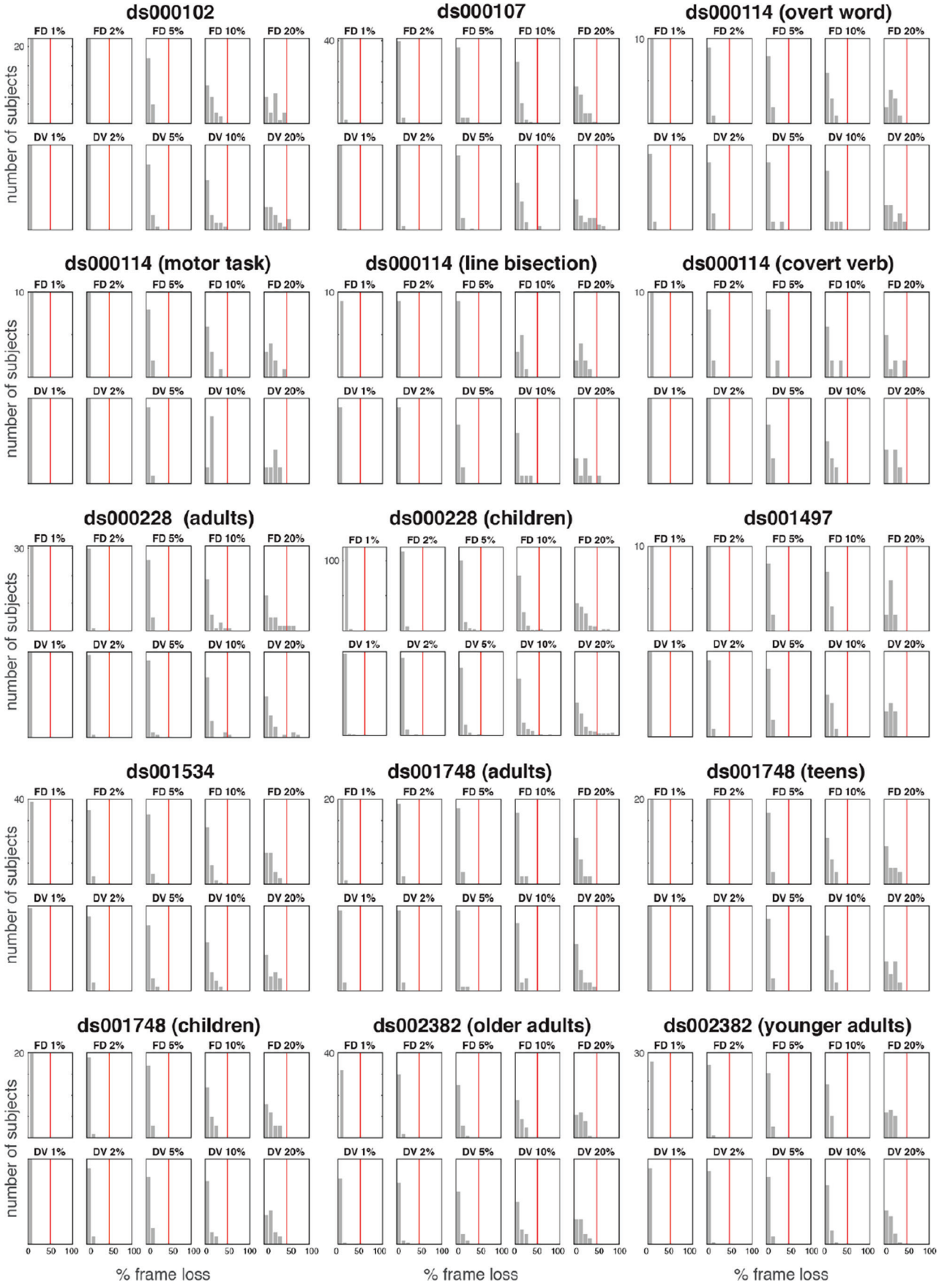

Fig. S3. Frameloss histograms for all 15 datasets and for all FD and DVARS thresholds. Red line indicates 50% frame loss for reference. See also Figure 2 in the main text.

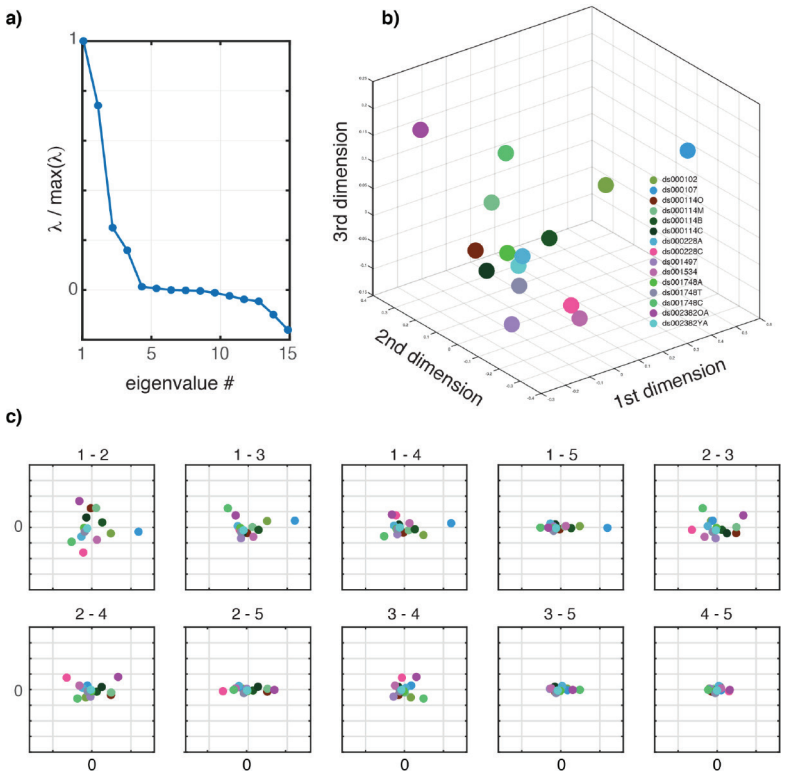

Fig. S4. Scree plot and additional MDS results. **(a)** Eigenvalue magnitude normalized to that of the largest eigenvalue returned by a multidimensional scaling of the 15 Dice overlap matrices shown in Figure 4 in the main text. Falloff suggests the dimensionality of the distance data is about five. The presence of negative eigenvalues indicates the dataspace is non-Euclidean. **(b)** Plot of first three eigen-dimensions reveals no obvious pattern in the data. **(c)** Projections of all possible pairs of the first five eigen-dimensions. The dataset ds000107 is segregated in some of the plots, but no other organization is apparent.

Table S1. Some prior evaluations of motion correction in task-based fMRI

Study	# Data sets	Task(s)	*N*	Motion correction approaches	Outcome measures
Current study	8	See Table 1	405	RP6, RP24, rWLS, WDS, untrained ICA, FD- and DVAR-based frame censoring	Whole-brain and ROI maximum *t*, mean effect size, Dice
Diedrichsen and Shadmehr (2005)	1	Hand controlled cursor targeting	15	Robust weighted least squares	*t*-Score, number of suprathreshold voxels
Hoffmann et al. (2015)	1	Auditory (SPM MoAE data contaminated with clinically relevant motion)	1	Rigid body alignment in SPM, FSL, AFNI, or AIR	Count of suprathreshold voxels, specious identification of motion
Huang et al. (2008)	1	Reading aloud	13	Linear interpolation over contaminated volumes	Significant voxel counts
Johnstone et al. (2006)	1	Go/no-go, N-back	33	Rigid body registration with versus without including six realignment parameters as covariates	Whole-brain maximum *t* and cluster extent
Kay et al. (2013)	11**	Visual	1–3	Custom (GLMDenoise), ICA	Cross validation
Kochiyama et al. (2005)	1	Finger tapping	4	ICA	ROC true versus false-positive fraction
Lemmin et al. (2010)	1	Arm motion	7	Model-based motion estimation, RP6, rWLS	Reduction in ventricle activation
Liao et al. (2006)	1	Motor	10	Custom ICA	Standard deviation reduction in activated voxels
Mayer et al. (2019)	2*	AX continuous performance task, multimodal attention	110	RP12, RP24, untrained ICA, trained ICA	Percent change of true and false activation
Middlebrooks et al. (2017)	1	Motor task, language task	12	Realignment, DVARS scrubbing, trained ICA, ICA+scrubbing	Increased *z*-scores in areas of expected activation
Oakes et al. (2005)	1	Go/No-go, N-back	40	Rigid body registration in AFNI, AIR, BrainVoyager, FSL, or SPM	Maximum *t* and cluster extent
Siegel et al. (2014)	4*	String matching, rule switching, posner task	88	Motion estimates as nuisance regressors versus FD-thresholded frame censoring	Change in β estimate, error variance, or *t*-score
Tierney et al. (2016)	1	Sentence comprehension and generation	42	Custom algorithm (FIACH) compared with RP6, RP24, FD frame censoring, rWLS, tCompCorr	ROI restricted *t*-values and cluster extent
Tohka et al. (2008)	2***	Tone counting, weather prediction	32	Trained ICA	Change in Z-scores
Wilke and Baldeweg (2019)	3	Verb generation, hand motor task, language task	84	Custom algorithm	*t*-Value, # activated voxels, SNR
Xu et al. (2014)	1	Overt speech	18	Custom ICA	PET cross validation

* Same scanner, multiple cohorts

** 11 task variations run on a single cohort

*** A training and a test cohort on the same scanner/tasks

Table S2. Acquisition details for datasets analyzed

Dataset	Reference	Scanner	Field strength (T)	TR (s)	Voxel size (mm)
ds000102	Kelly et al. (2008)	Siemens Allegra	3	2	3 × 3 × 4
ds000107	Duncan et al. (2009)	Siemens Avanto	1.5	3	3 × 3 × 3
ds000114	Gorgolewski et al. (2013)	GE Signa HDxt	1.5	2.5	4 × 4 × 4
ds000228	Richardson et al. (2018)	Siemens Tim Trio	3	2	3 × 3 × 3
ds001497	Lewis-Peacock and Postle (2008)	GE Signa VH/I	3	2	3.75 × 3.75 × 4
ds001534	Courtney et al. (2018)	Philips Intera Achieva	2.5	2.5	3 × 3 × 3
ds001748	Fynes-Clinton et al. (2019)	Siemens Magnetom Trio	3	3	2.5 × 2.5 × 2.5
ds002382	Rogers et al. (2020)	Siemens Prisma	3	3.07	2 × 2 × 2

Table S3. DVARS and FD values at percent frameloss

	FD					DVARS				
Dataset	1%	2%	5%	10%	20%	1%	2%	5%	10%	20%
ds000102	22.1	20.2	18.5	17.5	16.9	0.58	0.41	0.25	0.20	0.16
ds000107	32.3	29.5	26.8	25.3	24.2	0.52	0.41	0.27	0.20	0.13
ds000114 bisection	35.4	30.8	25.9	23.1	20.8	0.84	0.68	0.46	0.33	0.24
ds000114 covert verb	25.5	23.8	22.1	20.6	19.0	0.47	0.39	0.30	0.25	0.21
ds000114 lips motor	31.1	28.1	24.3	22.1	20.1	0.73	0.57	0.41	0.31	0.24
ds000114 overt word	33.7	31.8	28.6	26.3	23.7	0.57	0.44	0.35	0.28	0.21
ds000228 adults	30.1	25.3	20.9	17.4	15.1	0.99	0.77	0.55	0.41	0.33
ds000228 children	72.3	59.1	42.1	31.6	23.5	3.50	2.50	1.53	1.01	0.61
ds001497	284.6	251.3	219.4	198.3	179.6	0.60	0.45	0.33	0.24	0.18
ds001534	28.2	24.1	20.7	18.8	17.1	0.59	0.45	0.32	0.24	0.18
ds001748 adults	31.0	27.4	22.0	19.2	16.4	0.63	0.50	0.32	0.22	0.15
ds001748 children	58.0	50.5	38.4	30.0	23.5	1.55	1.19	0.81	0.57	0.37
ds001748 teen	58.7	50.1	35.9	28.6	23.7	1.07	0.76	0.48	0.33	0.22
ds002382 older adult	463.4	410.1	343.6	300.8	258.0	1.38	1.16	0.89	0.72	0.56
ds002382 young adult	382.0	324.3	258.8	222.1	193.2	1.16	0.85	0.55	0.38	0.27

Additional Dataset Details

Details of datasets used in the study are summarized below, including challenges or irregularities we encountered during analysis.

Accession Number: ds000102

Publication: Kelly et al. (2008)

Task: Slow event-related Eriksen flanker

Task details: Participants used one of two buttons to indicate the direction of a central arrow in an array of five arrows. In congruent trials, the flanking arrows pointed in the same direction as the central arrow; in more demanding incongruent trials, the flanking arrows pointed in the opposite direction.

Acquisition: Siemens Allegra 3.0 T (TR = 2000 ms; TE = 30 ms; flip angle = 80, 40 slices, matrix = 64 × 64; FOV = 192 mm; acquisition voxel size = 3 × 3 × 4 mm).

Number of subjects: 26; Age: 22–50 years (mean 32)

Contrast evaluated: incongruent correct > congruent correct

ROI: Neuroquery search term "incongruent task" and thresholded $Z > 3$

Notes: Functional data include a pronounced periodic artifact that appears unrelated to motion.

Accession Number: ds000107

Publication: Duncan et al. (2009)

Task: One-back

Task details: A one-back task was used with four categories of visual stimuli: written words, pictures of common objects, scrambled pictures of the same objects, and consonant letter strings. Subjects were instructed to press a button if the stimulus was identical to the preceding stimulus (12.5% of the stimuli were targets). Each block consisted of 16 trials from a single category presented one every second. A trial began with a 650 ms fixation cross, followed by the stimulus for 350 ms.

Acquisition: Siemens Avanto 1.5 T. The functional data were acquired with a gradient-echo EPI sequence (TR = 3000 ms; TE = 50 ms; FOV = 192 × 192; matrix = 64 × 64, voxel size = 3 × 3 × 3 mm).

Number of subjects: 45 (23 Male); Age: 19–38 years (mean 25)

Contrast evaluated: Words > 0

ROI: 5 mm sphere at [−42 −62 −16] (left ventral occipital–temporal cortex). Coordinates are from Table 2 of Duncan et al. (2009).

Notes: Data from six subjects were excluded: two because of corrupt data, three due to data modeling errors that could not be corrected, and one due to an incompatible contrast definition. Removed spaces from event names in BIDS .tsv files.

Accession Number: ds000114

Publication: Gorgolewski et al. (2013)

Tasks: (i) lip movement, (ii) covert verb generation, (iii) overt word generation, (iv) line bisection. All tasks used the same subjects.

Task details: Lip movement: Lip poaching (15 s) interleaved with fixation at a cross (15 s). Covert verb generation: Subjects instructed to think of a verb following presentation of a random noun for 1 s. Overt word generation: Repeat words aloud presented via headphones; 30 s task / 30 s rest repeated six times. Line bisection: Judge by button press if a horizontal line was bisected exactly in the middle (landmark) or if a horizontal line was crossed or not crossed (detection). Randomized presentation of six correct and four incorrect lines (525 ms presentation / 1100 ms response) in eight blocks.

Acquisition: GE Signa HDxt 1.5 T scanner with an 8-channel phased-array head coil. FOV = 256 × 256 mm, voxel size = 4 × 4 × 4 mm, slice thickness 4 mm, 30 slices per volume, interleaved slices order, acquisition matrix 64 × 64, flip angle = 90, TE = 50 ms, TR = 2.5 s, except for overt word repetition in which sparse sampling was used (TR = 5 s, "real TR"—which we assumed meant "acquisition time/TA" = 2.5 s). Subjects were scanned twice, either two or three days apart.

Number of subjects: 10 (4 Male); age: 50–58 years (median 52.5).

Contrast evaluated: (i) Motor task: lip > hand+foot, (ii) Covert verb: task > 0, (iii) Overt word: task > 0, (iv) Line bisection: landmark > detection

ROI: (i) 5 mm spheres in bilateral motor cortex ([−56, −6, 26] and [62, 0, 28]). Coordinates taken from (Pulvermüller et al. 2006), (ii) Covert verb: Broca's Area (BA 44 + BA 45 (left hemisphere only) from Anatomy Toolbox), (iii) Overt Word: same as lip motor task, (iv) Line bisection: lateral visual cortex (Anatomy Toolbox hoc3v+doc4v+hoc4lp).

Notes: The motor data included finger and foot tapping tasks that were not used because a mixture of left- and right-handed activation precluded straightforward second-level modeling. Subject 10 was excluded from the line bisection analysis as first level maps suggest the subject misunderstood task instructions (activation is left/right reversed).

Accession Number: ds000228

Publication: Richardson et al. (2018)

Task: Film viewing

Task details: Subjects viewed a 5.6-minute animated film with scenes classified as presenting either "pain" or "theory of mind" events.

Acquisition: 3T Siemens Tim Trio using a standard Siemens 32-channel head coil. T1-weighted structural images were collected in 176 interleaved sagittal slices with 1 mm isotropic voxels (GRAPPA parallel imaging, acceleration factor of 3; adult coil: FOV: 256 mm; kid coils: FOV: 192 mm). Functional data were collected with a gradient-echo EPI sequence in 32 interleaved near-axial slices aligned with the anterior/posterior commissure and covering the whole brain (EPI factor: 64; TR: 2 s, TE: 30 ms, flip angle: 90). Voxel size: Adults 3.13 × 3.13 × 3.13 mm; children either 3 × 3 × 3 mm or 3.13 × 3.13 × 3.13 mm.

Number of subjects: Adults: 33 (20 female); age: 18–39 years (mean: 24.8). Children: 123 (64 female); age: 3.5–12 years; mean: 6.7).

Contrast evaluated: pain > theory of mind

ROI: Regions listed in Supplementary Table 2 of Richardson et al. (2018), modeled as a collection of 5 mm spheres.

Notes: There were insufficient sessions in this data for test−retest evaluation. Event files were missing from the original OpenNeuro listing and were added manually using information provided in the description. The event timing provided was converted to seconds from scans using the TR information.

Accession Number: ds001497

Publication: Lewis-Peacock and Postle (2008)

Task: Stimulus judgment / memory

Task details: Subjects viewed a total of 90 stimuli drawn from three categories: 30 famous people, 30 famous locations, and 30 common objects. They indicated (on a four-point Likert scale, using a stimulus–response box) how much they liked the celebrity, how much they would like to visit the location, or how often they encountered the object in everyday life.

Acquisition: GE Signa VH/I 3T scanner. T1 (30 axial slices, 0.9375 × 0.9375 × 4 mm). Functional images: gradient-echo echo-planar (TR = 2000 ms; TE = 50 ms; 64 × 64 matrix coplanar with the T1 acquisition, voxel size = 3.75 × 3.75 × 4 mm).

Number of subjects: 10 (7 male); age: 19–32 years.

Contrast evaluated: Face > 0

ROI: Bilateral fusiform face area (Neuroquery search term "FFA" and thresholded Z > 3)

Notes: There were a total of six sessions in the data which were split into even and odd sessions for test–retest evaluation. The data on OpenNeuro is only the "LTM" portion of the experiment. Data from a working memory task described in the associated publication are not included.

Accession Number: ds001534

Publication: Courtney et al. (2018)

Task: Food images paired with textural calorie content

Task details: Participants first viewed images of food paired with an accompanying image number ("foodimage"), and subsequently viewed these same food images paired with the corresponding calorie information ("calorieimage"). The presentation sequence of food images and jittered fixation trials were pseudo-randomized.

Acquisition: Philips Intera Achieva scanner. Anatomical images were acquired using gradient-echo sequence (TR = 9.9 ms; TE = 4.6 ms; flip angle = 8; 1 × 1 × 1 mm voxels). Functional images were collected using T2* fast field echo (TR = 2.5 seconds, TE = 35 ms, flip angle = 90, voxel size = 3 × 3 × 3 mm).

Number of subjects: 50 (50 M); age: 18–22 years (mean 19.7).

Contrast evaluated: labeled image > not labeled

ROI: Bilateral inferior parietal cortex (Anatomy Toolbox IPC_PF + IPC_PFcm + IPC_PFm + IPC_PFop + IPC_PFt + PIC_PGa + IPC + PGp).

Notes: Functional images in this data were scaled by 0.02 prior to processing.

Accession Number: ds001748

Publication: Fynes-Clinton et al. (2019)

Task: Memory retrieval including autobiographical, episodic, or semantic conditions

Task details: One of 25 images of everyday life events was presented for 4 s, followed by a retrieval cue screen for 8 seconds during which participants retrieve different long-term memories. The type of memory retrieval was manipulated by adjusting the response screen to cue the retrieval of either personal experience (AM), general knowledge and factual information (SM), or questions about the content of the cue images (EM).

Acquisition: 3T Siemens scanner equipped with a 32-channel head coil. Structural: 176 slices sagittal; 1 mm isotropic volume; TR = 4000 ms; TE = 2.89 ms; FOV = 256 mm. Functional: T2*-weighted echo-planar image pulse sequence (45 slices, 2.5 mm slice thickness; voxel size = 2.5 × 2.5 × 2.5 mm, TR = 3000 ms; TE = 30 ms; FOV = 190 mm; flip angle = 90).

Number of subjects: 62 (32M); age: 10–35 years (see Notes).

Contrast evaluated: task > control

ROI: Inferior frontal gyrus (Neuroquery search term "IFG" thresholded Z > 3)

Notes: Data comprised three cohorts: children (10–12; *n* = 21), adolescents (14–16; *n* = 20), and young adults (20–35; *n* = 22) that were analyzed separately. There were insufficient data for test–retest evaluation. The tsv file for child-20 contains a typo with "semantic" mislabeled as "semanti" and autobio.tsv is empty for child-13. These subjects were excluded.

Accession Number: ds002382

Publication: Rogers et al. (2020)

Task: Speech comprehension in noise

Task details: Subjects were presented auditory stimuli via MR-compatible headphones consisting of words (monosyllabic consonant–vowel–consonant), silence, and noise (single-channel noise vocoded words) in two sessions of passive listening and two sessions of word repetition in which participants were asked to repeat aloud the presented word. Responses in the repeat condition were recorded and scored as either correct or incorrect.

Acquisition: MRI data were acquired using a Siemens Prisma scanner (Siemens Medical Systems) at 3 T equipped with a 32-channel head coil. Scan sequences began with a T1-weighted structural volume using an MPRAGE sequence (TR = 2.4 s, TE = 2.2 ms, flip angle = 8, 300 × 320 matrix, voxel size = 0.8 mm isotropic). Functional images were acquired using a multiband echo-planar imaging sequence (TR = 3.07 s, TA = 0.770 s, TE = 37 ms, flip angle = 37, voxel size = 2 × 2 × 2 mm, multiband factor = 8).

Number of subjects: Young adults: *n* = 29 (19 female); age: 19–30 years (mean = 23.8). Older adults: *n* = 32 (17 female); age 65–81 years (mean = 71.0).

Contrast evaluated: repeat word > noise

ROI: 5 mm spheres in bilateral motor cortex ([−56, −6, 26] and [62, 0, 28]). Coordinates taken from Pulvermüller et al. (2006)

Notes: Young adults and older adults were analyzed separately.

REFERENCES

1. Courtney AL, PeConga EK, Wagner DD, Rapuano KM. Calorie information and dieting status modulate reward and control activation during the evaluation of food images. PLoS One. 2018 Nov 2;13(11):e0204744.
2. Diedrichsen J, Shadmehr R. Detecting and adjusting for artifacts in fMRI time series data. Neuroimage. 2005 Sep;27(3):624–34.
3. Duncan KJ, Pattamadilok C, Knierim I, Devlin JT. Consistency and variability in functional localisers. Neuroimage. 2009;46(4):1018–26.
4. Fynes-Clinton S, Marstaller L, Burianová H. Differentiation of functional networks during long-term memory retrieval in children and adolescents. Neuroimage. 2019 May 1;191:93–103.
5. Gorgolewski KJ, Storkey A, Bastin ME, Whittle IR, Wardlaw JM, Pernet CR. A test-retest functional MRI dataset for motor, language and spatial attention functions [Internet]. GigaScience Database; 2013. Available from: http://gigadb.org/dataset/100051
6. Hoffmann M, Carpenter TA, Williams GB, Sawiak SJ. A survey of patient motion in disorders of consciousness and optimization of its retrospective correction. Magn Reson Imaging. 2015 Apr;33(3):346–50.
7. Huang J, Francis AP, Carr TH. Studying overt word reading and speech production with event-related fMRI: a method for detecting, assessing, and correcting articulation-induced signal changes and for measuring onset time and duration of articulation. Brain Lang. 2008 Jan;104(1):10–23.
8. Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, Davidson RJ, et al. Motion correction and the use of motion covariates in multiple-subject fMRI analysis. Hum Brain Mapp. 2006 Oct;27(10):779–88.
9. Kay KN, Rokem A, Winawer J, Dougherty RF, Wandell BA. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front Neurosci. 2013 Dec 17;7:247.
10. Kelly AMC, Uddin LQ, Biswal BB, Castellanos FX, Milham MP. Competition between functional brain networks mediates behavioral variability. Neuroimage. 2008 Jan 1;39(1):527–37.
11. Kochiyama T, Morita T, Okada T, Yonekura Y, Matsumura M, Sadato N. Removing the effects of task-related motion using independent-component analysis. Neuroimage. 2005 Apr 15;25(3):802–14.
12. Lemmin T, Ganesh G, Gassert R, Burdet E, Kawato M, Haruno M. Model-based attenuation of movement artifacts in fMRI. J Neurosci Methods. 2010 Sep 30;192(1):58–69.
13. Lewis-Peacock JA, Postle BR. Temporary activation of long-term memory supports working memory. J Neurosci. 2008 Aug 27;28(35):8765–71.
14. Liao R, McKeown MJ, Krolik JL. Isolation and minimization of head motion-induced signal variations in fMRI data using independent component analysis. Magn Reson Med. 2006 Jun;55(6):1396–413.
15. Mayer AR, Ling JM, Dodd AB, Shaff NA, Wertz CJ, Hanlon FM. A comparison of denoising pipelines in high temporal resolution task-based functional magnetic resonance imaging data. Hum Brain Mapp. 2019 Sep;40(13):3843–59.
16. Middlebrooks EH, Frost CJ, Tuna IS, Schmalfuss IM, Rahman M, Old Crow A. Reduction of motion artifacts and noise using independent component analysis in task-based functional MRI for preoperative planning in patients with brain tumor. AJNR Am J Neuroradiol. 2017 Feb;38(2):336–42.
17. Oakes TR, Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, et al. Comparison of fMRI motion correction software tools. Neuroimage. 2005 Nov 15;28(3):529–43.
18. Pulvermüller F, Huss M, Kherif F, Moscoso del Prado Martin F, Hauk O, Shtyrov Y. Motor cortex maps articulatory features of speech sounds. Proc Natl Acad Sci U S A. 2006 May 16;103(20):7865–70.
19. Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. Development of the social brain from age three to twelve years. Nat Commun. 2018 Mar 12;9(1):1027.
20. Rogers CS, Jones MS, McConkey S, Spehar B, Van Engen KJ, Sommers MS, et al. Age-related differences in auditory cortex activity during spoken word recognition. Neurobiol Lang. 2020 Oct;1(4):452–73.
21. Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL, et al. Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. Hum Brain Mapp. 2014;35:1981–96.
22. Tierney TM, Weiss-Croft LJ, Centeno M, Shamshiri EA, Perani S, Baldeweg T, et al. FIACH: a biophysical model for automatic retrospective noise control in fMRI. Neuroimage. 2016 Jan 1;124(Pt A):1009–20.
23. Tohka J, Foerde K, Aron AR, Tom SM, Toga AW, Poldrack RA. Automatic independent component labeling for artifact removal in fMRI. Neuroimage. 2008 Feb 1;39(3):1227–45.
24. Wilke M, Baldeweg T. A multidimensional artefact-reduction approach to increase robustness of first-level fMRI analyses: censoring vs. interpolating. J Neurosci Methods. 2019 Apr 15;318:56–68.
25. Xu Y, Tong Y, Liu S, Chow HM, AbdulSabur NY, Mattay GS, et al. Denoising the speaking brain: toward a robust technique for correcting artifact-contaminated fMRI data under severe motion. Neuroimage. 2014 Dec;103:33–47.
26. Zar JH. Biostatistical Analysis. 5th ed. Prentice-Hall/Pearson. 2010.

