Motion denoising of multiband resting state functional connectivity MRI data: An improved volume censoring method

John C. Williams [a, b]

Jared X. Van Snellenberg* [a, b, c, d, e]

[a]Department of Psychiatry and Behavioral Health, Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, USA

[b]Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY, USA

[c]Department of Psychology, Stony Brook University, Stony Brook, NY, USA

[d]Division of Translational Imaging, New York State Psychiatric Institute, New York, NY, USA

[e]Department of Psychiatry, Columbia University Medical Center, New York, NY, USA

CORRESPONDING AUTHOR: Jared X. Van Snellenberg, 101 Nicolls Rd, Health Sciences Center T10-087J, Stony Brook, NY 11794

KEYWORDS: Resting State Functional Connectivity, fMRI, Motion, Volume Censoring

Abstract

The study of resting state functional connectivity (RSFC) using functional MRI scans has rapidly become one of the most promising and widely used techniques for investigations of human brain function in both healthy and psychiatric populations. Two critical recent developments in this field are 1) the increasing use of simultaneous multi-slice fMRI (multiband) acceleration techniques, which dramatically improves the spatial and temporal resolution of fMRI data, and 2) the recognition that participant motion is a critical confound in RSFC studies, which requires careful denoising in order to obtain valid results. However, motion artifact denoising techniques were not developed with the temporal resolution of multiband fMRI in mind, which results in the capture of high-frequency respiration-related motion of participants during scanning. This respiration-related motion appears to negatively impact the performance of existing volume censoring approaches. Using publicly available multiband RSFC data from the Human Connectome Project, we developed a new volume censoring motion correction approach that addresses respiration-related motion separately from other sources of motion, and outperforms one of the most widely used denoising pipelines. We further show that the assumptions underlying some of the most commonly employed metrics for evaluating motion denoising pipelines (testing for significant differences in RSFC correlations between high- and low-motion participants, and so-called "QC-FC" based methods) are invalid. Specifically, the number of significant RSFC correlations between high- and low-motion groups is dramatically reduced by exclusion of participants exhibiting substance use or who have a family history of psychiatric or neurological disorder, indicating that individual differences in unmeasured third variables contribute to both higher motion and true differences in RSFC correlations, thereby invalidating this widely used metric, which assumes that no true differences in RSFC exist

between high- and low-motion groups. Finally, we develop and present an empirical basis for selecting volume censoring thresholds in any multiband RSFC dataset, which are widely used in the field but have had an exclusively heuristic basis prior to this work. These findings thus have three major impacts: first, to present a substantively improved pipeline for motion denoising of multiband RSFC data; second, to raise concerns about a key metric used to evaluate motion denoising for RSFC data more generally; and third, to provide investigators with an empirically-grounded estimate of the optimal volume censoring threshold to employ for any dataset.

Introduction

Over the past decade and a half, the use of functional Magnetic Resonance Imaging (fMRI) to investigate resting state functional connectivity (RSFC) has become the dominant approach to studying the connectome of the human brain, a key priority of the National Institute of Mental Health for achieving their strategic objective to define the mechanisms of complex behaviors[1]. However, in recent years it has been increasingly recognized that participant motion represents a serious confound for investigations using RSFC, and that aggressive steps must be taken to minimize the impact of this confound[2-11]. The success of this ongoing effort will be crucial to the utility of RSFC as a tool for human connectomics in both health and disease.

In recent years, simultaneous multi-slice acceleration of fMRI sequences (multiband fMRI)[12,13] has gained prominence in RSFC research, and has been adopted by several large-scale studies of human brain function, including the Human Connectome Project (HCP)[13], the UK Biobank study[14], and the Adolescent Brain Cognitive Development (ABCD) study[15-18]. Multiband acceleration provides substantial improvements in both spatial and temporal resolution of both RSFC and task-based fMRI data. Critically, however, the performance of existing RSFC motion denoising methods for multiband data has not been extensively evaluated (but see [19]), although recent work demonstrates that the fast sampling rate (the inverse of repetition time; TR) of multiband data necessitates the development of new approaches for handling estimates of participant motion, such as framewise displacement (FD)[20,21].

Early in our work with multiband RSFC data, we noted that estimated participant motion parameters (MPs) universally exhibit a high-frequency signal component that varies in frequency between participants, typically falling between 0.2 and 0.4 Hz (**Figure 1**). Other authors have recently demonstrated that these signals result from head motion due to participant respiration[20-

[22], although a small portion arises from "false" pseudo-motion in the phase-encode direction that occurs as a result of tissue changes due to lung expansion impacting the B_0 field[20]. In many participants, an elevation of average gray matter signal amplitude at the same frequency as the MP respiratory signal is also apparent (**Figure 1c**), suggesting that this motion effect also has a repeated, systematic, effect on fMRI Blood Oxygen Level-Dependent (BOLD) signal at high frequencies. Notably, any effect at these frequencies would be impossible to observe in higher TR single-band fMRI, because these sequences rarely achieve a sampling rate faster than 0.5 Hz (TR = 2 s). Due to the Nyquist limit, signals above 0.25 Hz would be aliased in these data. In contrast, the public HCP dataset was acquired at just under 1.4 Hz (TR = 0.72 s), allowing for characterization of signal components up to nearly 0.7 Hz.

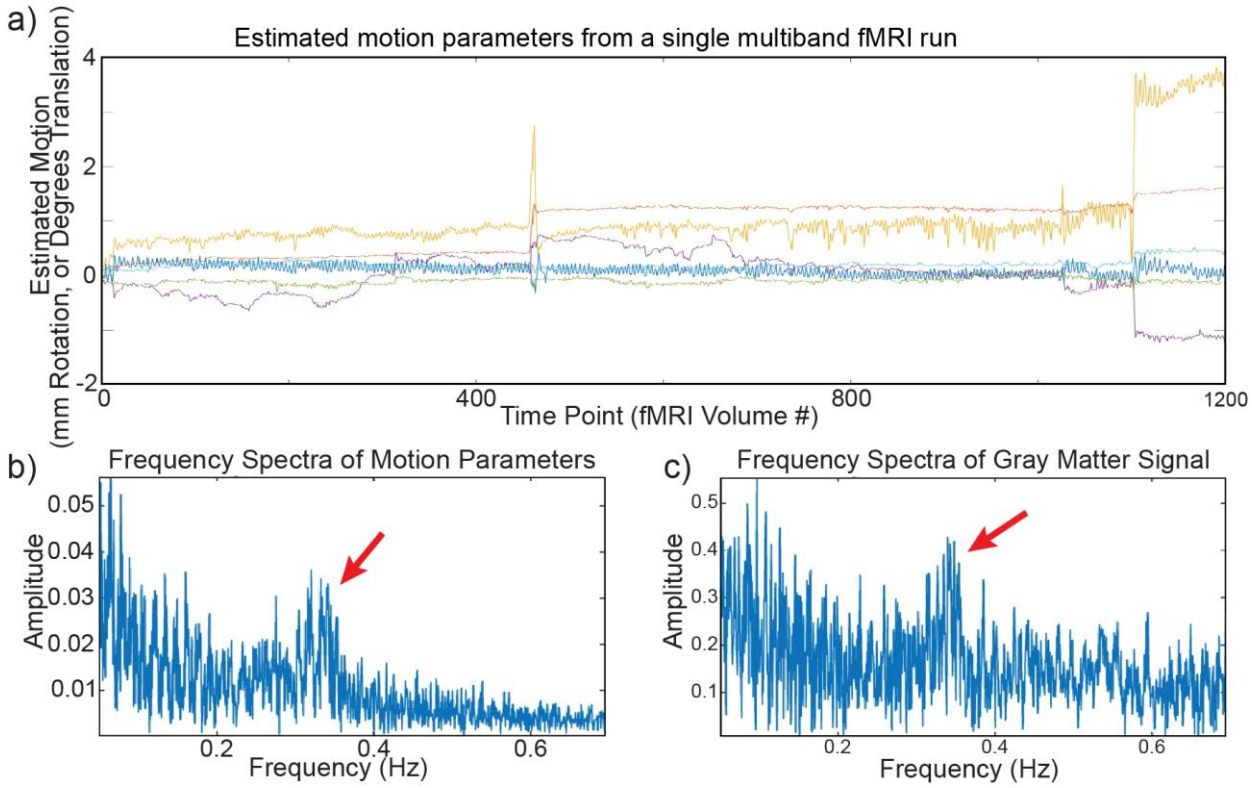

Figure 1. a) Estimated motion parameters (MPs) for a single RSFC scan from a participant in the Human Connectome Project (Subject ID 116524), which show a clear high-frequency oscillation throughout the run. **b)** Spectral frequency amplitudes of the MP traces, obtained via Fourier decomposition and summed over all 6 estimated MPs. **c)** Spectral frequency amplitudes of the mean signal extracted from all gray matter voxels after image preprocessing (see *Methods*). Red arrows indicate the elevation of spectral amplitude in both MPs and global signal thought to result from participant respiration.

We reasoned that because RSFC data is routinely band-pass filtered well below these frequencies (approximately 0.01 - 0.08 Hz), these respiration-related gray matter signal fluctuations likely have no impact on RSFC correlations calculated with filtered data. However, two of the most widely employed methods for motion correction in the field—volume censoring and spike regression[5,7]—are based on unfiltered estimates of subject motion (FD) and whole-brain signal change (temporal Derivative of timecourses, root-mean-squared VARiance over voxels; DVARS or DV)[4-6,23]. Thus, these denoising techniques would unnecessarily flag volumes affected by respiration-related motion and remove them from the dataset. Indeed, multiple other authors have concluded that employing a band-stop (notch) filter on MPs prior to calculating FD facilitates the identification of non-respiratory subject head motions[20,21], although these authors did not compare the effect of employing filter-based FD values for volume censoring to methods using unfiltered, or raw, FD values, nor did they employ similar methods on fMRI timeseries prior to calculation of DV. In contrast, our primary goal here is to determine optimal methods for addressing motion artifacts in multiband RSFC data, in order to provide clear recommendations to the field as to how multiband data should be processed in order to optimize the detection of true neural patterns of RSFC, as free as possible from the effect of head motion and related signal artifacts.

To this end, we hypothesized that a denoising approach that ignores motion due to respiration and censors only those volumes that exhibit irregular and sudden (and thus predominantly low frequency) motion might perform as well or better than existing methods. Consequently, we developed a volume censoring approach based on low-pass filtering at 0.2 Hz, which is just low enough to remove nearly all respiration-related motion while remaining at least double the upper frequency band of standard RSFC band-pass filters (which rarely exceed 0.1 Hz

for the upper end of the pass-band). We apply this filter to both MPs and to within-brain voxel timeseries prior to calculation of FD and DV, respectively, which we term LPF-FD and LPF-DV censoring, each with a threshold parameter, respectively termed Φ_F and Φ_D.

Further, we observed that both DV and LPF-DV values calculated for each participant, and even for each run within a participant, frequently exhibit large differences in central tendency, such that an apparently reasonable cutoff for one run could potentially remove all data from another. **Figure 2a-b** demonstrates this issue in four example runs from two subjects. Consequently, we also tested an adaptive thresholding method that fits a generalized extreme value distribution to the LPF-DV values within each run separately, and rejects the upper tail (i.e., outliers) of that distribution by setting Φ_D separately for each run, which we term GEV-DV censoring (see **Figure 2c**) with free parameter d_G.

Thus, below we evaluate the performance of each censoring approach (LPF-FD, LPF-DV, and GEV-DV) relative to standard volume censoring methods[5] that have performed well in recent evaluations of multiple motion denoising strategies[2,3], using the HCP S500 release[24]. We initially set out to employ a set of RSFC quality control (QC) metrics that have been employed in prior work[2-5], but we observed substantive issues with QC measures that depend on associations between QC measures (such as median FD) and the observed magnitude of RSFC correlations (e.g., so-called QC-FC correlations, or tests for significant RSFC differences between high- and low-motion groups). These issues, which we detail in the Results below, led us to adopt a new QC metric similar to graphical metrics employed in prior work[4,5]. This technique calculates the mean absolute change (MAC) in region of interest (ROI) pair correlations after volume censoring, over and above the change due to randomly removing an equivalent number of volumes, which we refer to as MAC-RSFC.

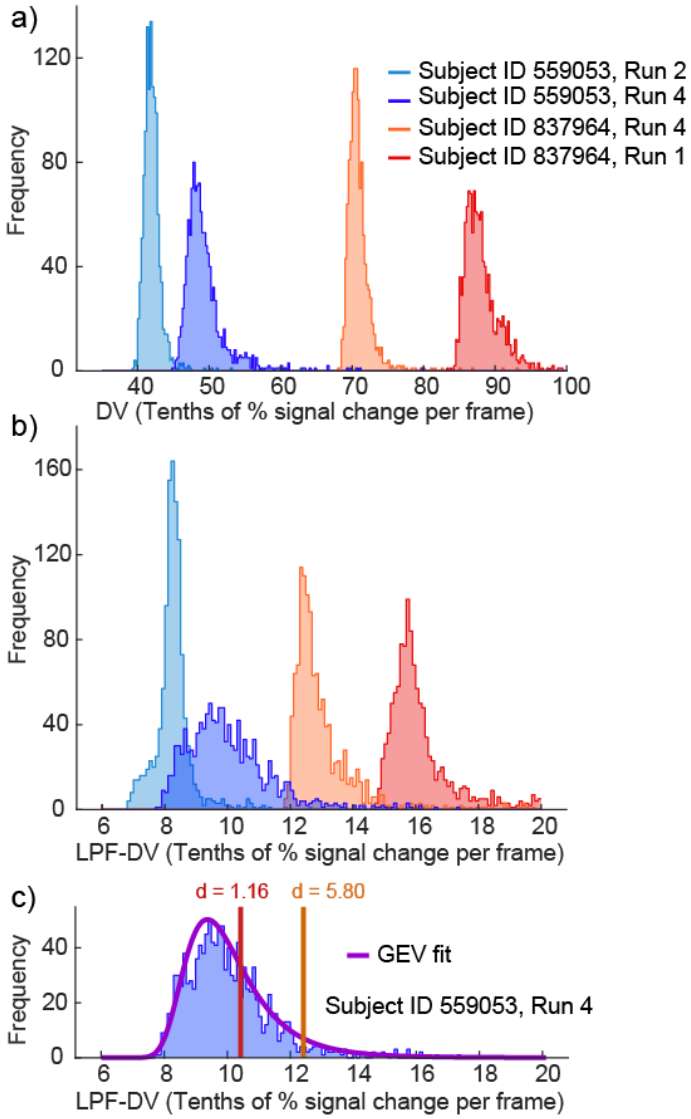

Figure 2. a) Histograms of framewise DV values for 4 runs, comprising 2 pairs of runs each from 2 subjects, demonstrating the large variability in central tendency of DV values between runs. **b)** Histograms of framewise LPF-DV for the same 4 runs as in A. **c)** Example generalized extreme value (GEV) distribution probability density function (PDF), determined from a maximum-likelihood fit, for one run (purple). Two adaptive censoring thresholds for this run are shown using (arbitrary unit) free parameter values d = 1.16 and d = 5.80.

As a final goal, we also sought to develop for the first time a quantitative, empirical basis for determining the optimal value of censoring thresholds for removing motion artifacts from RSFC data. That is, each volume censoring method requires the use of a free parameter that determines the aggressiveness of the censoring approach (e.g., two commonly employed parameters over the last several years are raw FD thresholds of 0.2 or 0.5 mm). However, to date

these thresholds have been relatively arbitrarily determined, based predominantly on visual inspection of post-processed data[5,6]. Instead, here we develop an approach that attempts to optimize the trade-off between reduction in motion-induced bias and the loss of power that both result from the removal of data due to censoring, using a statistically-driven metric we term *ΔMSE-RSFC* (actually the change in MSE/N; see *Results*, below). This approach uses a bias-variance decomposition of mean squared error (MSE) for estimators of RSFC correlations to approximate the *total* error in sample RSFC correlation estimates, including 1) motion-induced bias in sample mean correlations; 2) variance resulting from both true variation between sampled individuals and the unequal distribution of motion, and thus motion artifact, across individuals and runs; and 3) sampling error within each run. The resulting MSE estimate is then divided by the remaining post-censoring sample size to account for the increase in the width RSFC confidence intervals resulting from a loss of sample size (i.e., when participants are removed due to insufficient data remaining following censoring).

Thus, *ΔMSE-RSFC* seeks to provide a tangible minimization target for balancing the improvement in RSFC estimates through reduced motion-induced bias against the reduction in statistical power produced by a loss of observations (as frames are censored and removed from analysis). We then used a global optimization algorithm to simultaneously determine optimal parameters for both FD- and DV-based censoring in the HCP500 dataset by minimizing ΔMSE-RSFC. We repeated this analysis with and without global signal regression (GSR), and provide an analytically based method to generalize these results to fMRI datasets of nearly any size.

Results

Initial Comparison of Standard and LPF-Based Censoring

As an initial step to evaluate each censoring method, we sought to replicate a graphical approach to evaluating the efficacy of motion censoring employed in prior work[4,5], which shows changes in RSFC correlations across all ROI pairs (taken from a commonly used set of RSFC ROIs[25]) resulting from motion-targeted volume censoring (red dots; see **Figure 3**). In order to ensure that these changes are a result of removing high-motion volumes specifically, they are compared to RSFC changes resulting from removal of randomly selected volumes (black dots). To allow for a direct comparison between censoring methods in **Figure 3**, thresholds were selected such that an equivalent number of volumes were removed by each evaluated method (56.15% of volumes in analyses without GSR and 40.06% of volumes with GSR).

Standard censoring methods resulted in minimal changes in RSFC compared to random censoring of an equivalent number of frames within each run, consistent with our hypothesis that these methods would largely flag volumes impacted by respiratory motion for removal, and that removing these volumes would have little impact on RSFC correlations. In addition, standard

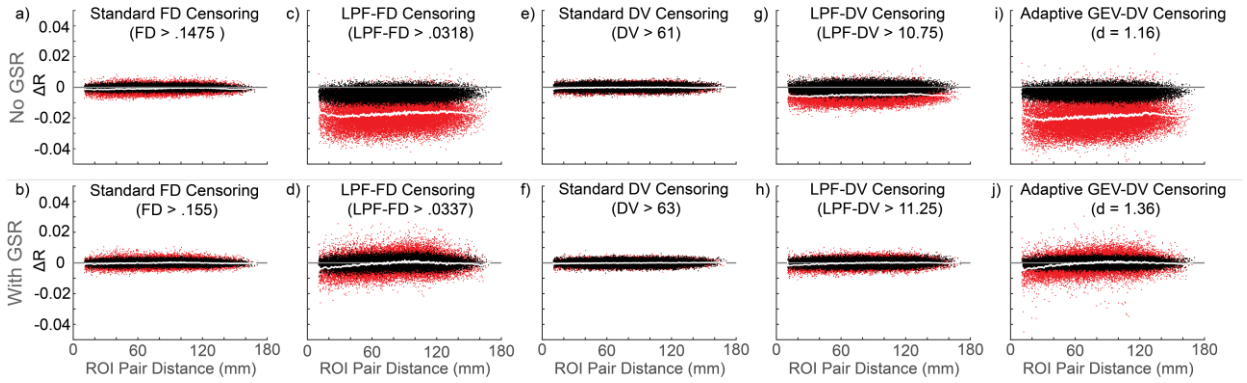

Figure 3. Changes in all pairwise ROI RSFC correlations resulting from volume censoring using raw FD **(a,b)** and raw DV **(e,f)**, LPF-FD **(c,d)** and LPF-DV **(g,h)**, and adaptive GEV-DV **(i,j)**, plotted against the Euclidean distance between ROIs on the X-axis. Red dots indicate the average change across the sample in RSFC correlation resulting from volume censoring for each ROI pair. Black dots indicate the average change resulting from random censoring of an equal number of volumes within each run. White lines indicate a sliding window of the mean of the red dots, Results are shown both without global signal regression (GSR) **(a,c,e,g,i)** and with GSR **(b,d,f,h,j)**.

methods were outperformed by LPF-based censoring, and GEV-DV censoring produced an even greater change in observed RSFC than LPF-DV censoring. Finally, in analyses without GSR, LPF-FD, LPF-DV, and GEV-DV censoring produced an overall downward shift in RSFC magnitude that was not observed in analyses employing GSR. Given that GSR has been shown to globally reduce the magnitude of RSFC correlations[2,26], this suggests that LPF-based methods may be producing some of the same effect as GSR in minimizing the impact of motion on overall RSFC correlation magnitude, and that this effect is consistent with a reduction of "Type 2" motion artifact effects on RSFC[6]. Similar results were obtained using an alternative method of comparing the differential performance of volume censoring methods, using one method as a baseline for evaluating the change in RSFC correlations produced as a result of censoring additional volumes targeted by another method (**Figure S1;** see *Supplementary Results*).

Comparison of High- to Low-Motion Participants After Censoring

Our next step was to look for significant differences in RSFC correlations between high- and low-motion participants before and after volume censoring, following methods employed in multiple prior studies[3,5,7,9]. The logic of this approach rests on the assumption that, as a group, high- and low-motion participants differ only in the amount of in-scanner head motion, and so any significant differences in RSFC must result from unmitigated motion artifacts rather than true differences in RSFC. However, when we employed this approach with LPF-FD censored data we observed null hypothesis rejection rates approximately two orders of magnitude higher than the nominal false positive rate, leading us to hypothesize that these observed differences may not, in fact, reflect false positives. Rather, there may be true RSFC differences between

high- and low-motion participants that are not strictly due to in-scanner motion (i.e., a "third-variable" effect).

For example, even if participants themselves are psychiatrically and neurologically in the normal range, a family history of these disorders, or a history of substance use, could produce *both* increased participant motion and true RSFC differences. **Figure 4** shows the observed proportion of RSFC correlations exceeding various uncorrected P value thresholds for both the full HCP 500 dataset and for a reduced dataset that excludes participants who had a parent with any psychiatric or neurological disorder (FH+ subjects), or who used any illicit substance or had a blood-alcohol level above 0.05 during the course of the study (SU+ subjects). As expected, FH+ and SU+ subjects had elevated median FD values (mean = 16.2) relative to FH- and SU- participants (mean = 14.7; $t_{490} = 3.06$; $P = 0.002$). **Figure 4** also shows 95% confidence intervals from a Monte Carlo simulation of the effect of removing an equal number of randomly selected participants who exhibited an equivalent amount of motion to FH+ and SU+ subjects, but who were themselves FH- and SU- (see *Methods*).

We observed that removing FH+ and SU+ participants causes a significantly greater reduction in observed null hypothesis rejection rates than would be expected simply by removing an equivalent number of FH- and SU- participants who exhibit similar levels of motion. Thus, these findings are consistent with the "third-variable" effect hypothesized above, such that FH+ and SU+ participants exhibit *both* true differences in RSFC relative to FH- and SU- individuals, *and* elevated in-scanner motion, thereby producing a *true association* between RSFC and motion that is independent of motion-induced signal artifacts. These results are also consistent with previous findings that trait effects were still detectable in RSFC correlations in the HCP dataset, even after aggressive denoising methods were employed[27].

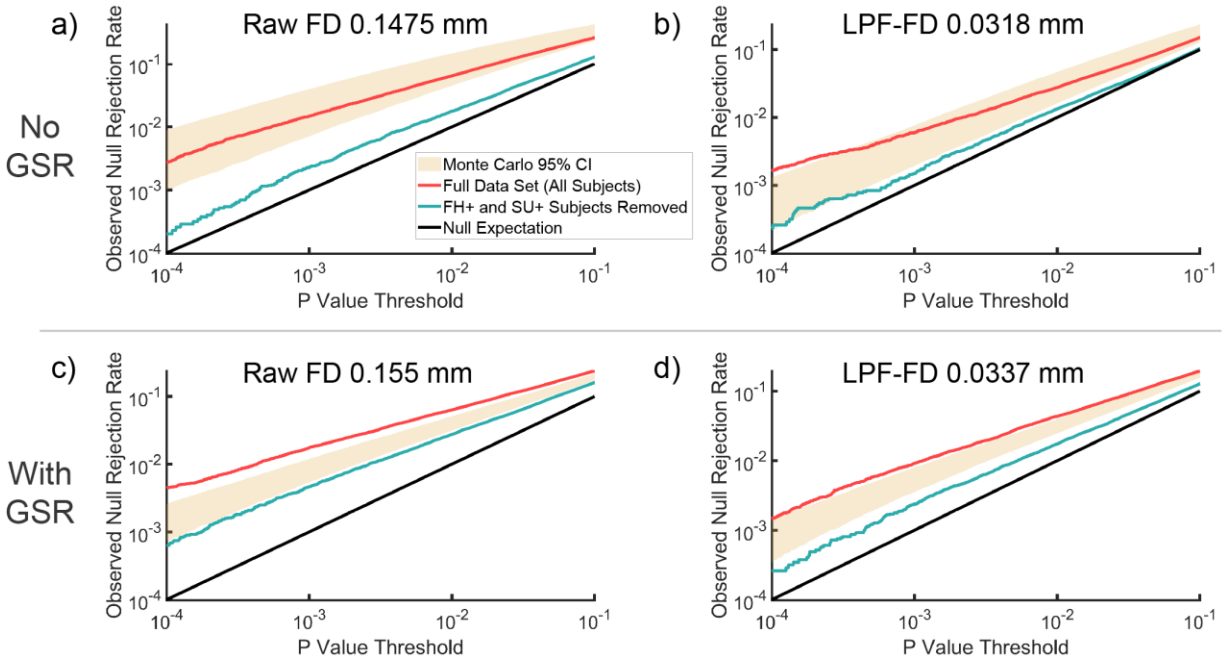

Figure 4. Observed proportion of RSFC correlations exceeding P value threshold (uncorrected for multiple comparisons). Red lines denote the proportion observed for the full dataset; blue-green lines denote the proportion observed in the dataset after removing subjects who tested positive for drug use or elevated blood alcohol content on the day of a scan session (SU+) or have a family history of a psychiatric or neurological disorder (FH+). 95% confidence intervals from 1,000 Monte Carlo resamplings of the SU-/FH- subset of the data with motion characteristics matched to the original dataset (matched on FD; see *Methods*) are shown in the beige area. Black lines represent the null expectation. Results are shown using both raw FD **(a,c)** and LPF-FD **(b,d)**, and without GSR (top) and with GSR (bottom).

These results argue strongly against using any method based on individual differences in the quantity of motion, such as testing for differences in RSFC correlations in high- versus low-motion participants, or measures that examine the correlation between participant motion and RSFC correlations in each ROI (commonly known as QC-FC correlations), which have been widely employed to date[2,3,5,6,8,19,28-32]. Even if one excludes participants with known "third variables" that influence both in-scanner motion and RSFC correlations, such as family history of psychiatric or neurological disorder and substance use as we show here, there is no guarantee that other third variables influencing motion and RSFC do not exist. Consequently, we would argue strongly against the use of such measures for evaluating the effectiveness of methods designed to reduce or eliminate the impact of participant motion on RSFC, unless it can be shown somehow that such third variables could not possibly exist in the data under examination

(e.g., with phantom or simulated data, or potentially within-subject comparisons that depend only on a manipulation that systematically effects within-scanner motion).

Comprehensive Evaluation of Standard and LPF-Based Censoring

As a result of the significant issues with measures depending on individual differences in motion (e.g., median FD values for each participant) described above, we sought to develop an alternative, quantitative metric to evaluate the efficacy of motion denoising efforts. Given that we view any method relying on the relationship between RSFC correlations and participant motion as potentially "contaminated" by *true* motion-RSFC relationships, we opted to focus directly on the magnitude of changes in RSFC correlations as a result of motion-targeted volume censoring (compared to random removal of an equivalent number of volumes within each run) within each subject, and averaged across subjects. Specifically, we evaluated the mean absolute value of the within-subject change in RSFC correlations relative to removing an equivalent number of volumes within each run, across all ROI pairs, and across all subjects in the sample. This provides a benchmark for the magnitude of systematic changes in RSFC correlation estimates within each subject produced by each method for motion-targeted censoring, relative to removing an equivalent number of frames within each run. Thus, observed between-method differences in this metric should be specifically associated with differences in targeting of BOLD signal fluctuations resulting from head motion, rather than any other source. We term this measure MAC-RSFC (Mean Absolute Change in Resting State Functional Connectivity).

A pervasive issue with volume censoring approaches to handling motion in RSFC (such as "scrubbing" and spike regression) has been a lack of clarity as to the appropriate threshold above which to censor RSFC volumes; i.e., the question as to how much motion is "too much."

In order to address this question, we tested a wide range of parameter values for both LPF and standard censoring methods, and employed the Stony Brook SeaWulf high performance computing cluster to conduct bootstrapping on the results at each parameter value, thereby allowing us to obtain confidence intervals across the full range of potential parameter values for all evaluated methods. As can be seen in **Figure 5,** which summarizes analyses comprising over 6.95×10^{15} partial correlations, both LPF-FD and LPF-DV produced larger magnitude changes in RSFC correlations than standard FD- and DV-based censoring, relative to random removal of volumes, regardless of how much data is removed by each method. That is, across the full range from minimal to maximal removal of data (approaching 0% and 100% of data removed), the LPF-based methods we propose here significantly (based on the non-overlap of 95% confidence intervals obtained from a bootstrapping procedure) outperformed standard censoring methods. In addition, GEV-DV outperformed LPF-DV, suggesting that, in line with our observations in **Figure 2**, **Figure 3g-j,** and **Figure S1i-l**, an adaptive thresholding method is preferable for handling the substantial differences in central tendency of DV measures across runs.

As noted above, other authors have recently suggested employing band-stop (notch) filters to separate respiration-related motion from other motion in fast-TR data such as the HCP dataset[20,21]. Although they did not directly compare the efficacy of a band-stop filter to other methods (nor did they propose applying such filters to voxelwise data prior to calculation of DV), we show that although these methods both outperform standard raw, unfiltered, approaches to FD-based censoring, they are both outperformed by the LPF methods presented here (**Figure S2)**. Thus, LPF-FD and GEV-DV volume censoring for multiband fMRI, at least in the HCP 500 dataset evaluated here, appears to be an across-the-board superior method of volume censoring, no matter how aggressive (or not) an investigator prefers to be in censoring motion-effected data.

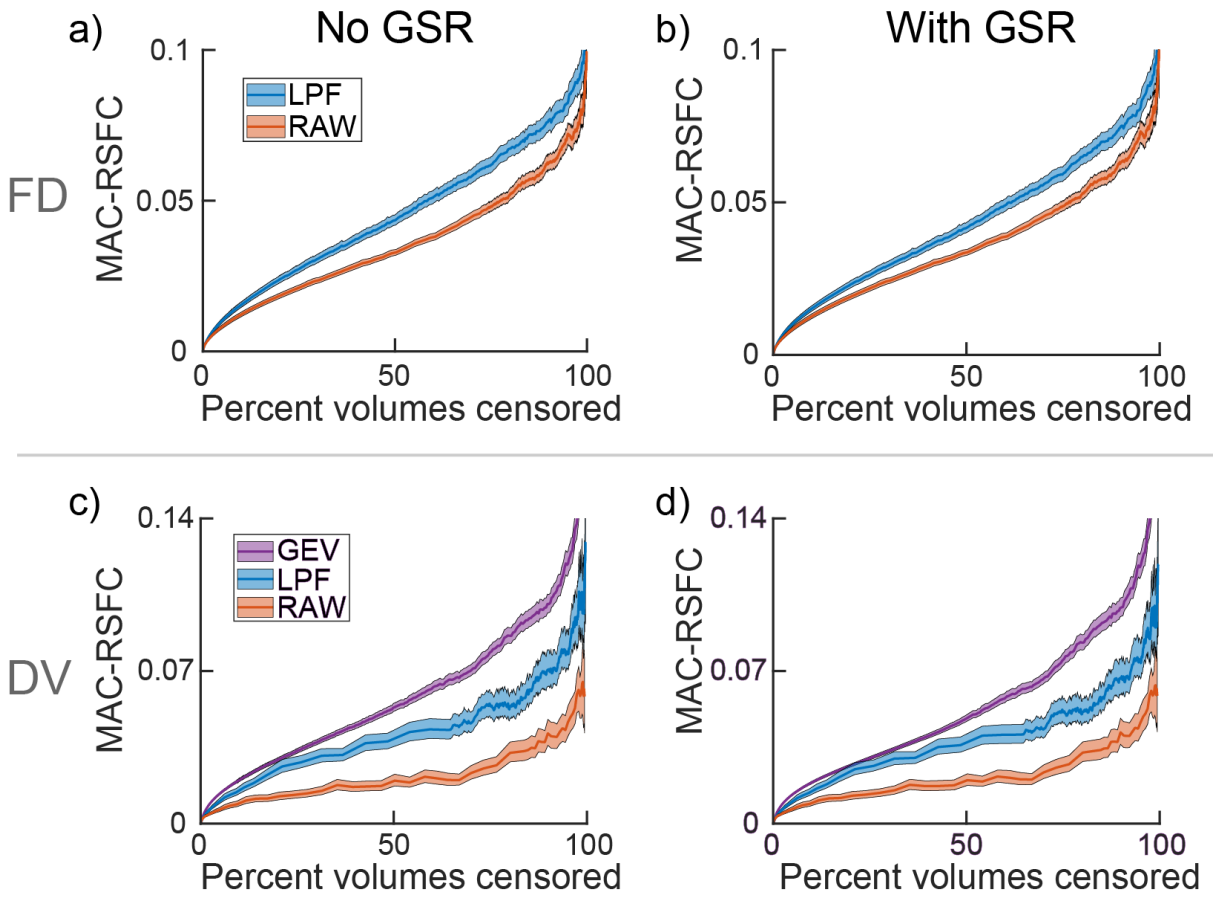

Figure 5. Mean Absolute Change in Resting State Functional Connectivity (MAC-RSFC) after volume censoring using standard FD and LPF-FD **(a,b)** and standard DV, LPF-DV, and GEV-DV **(c,d)** across a range of parameter values. 95% CIs on MAC-RSFC were estimated via bias corrected and accelerated bootstrapping with 10,000 bootstrap samples. Results are shown both without GSR **(a,c)** and with GSR **(b,d)**.

Optimization of Volume Censoring Thresholds: A Bias-Variance Decomposition Approach

It has been noted that, irrespective of the method used to target high-motion volumes for removal, use of increasingly strict thresholds for volume censoring results in continuous 'improvements' in data quality (i.e., continuous reduction in motion-induced bias in RSFC correlations)[5,6,33], an observation that is also clearly demonstrated in **Figure 5**, above. This raises a critical question for investigators as to how to balance the tradeoff between the benefits of additional denoising (i.e., selecting a censoring parameter that results in a more restrictive threshold), and the costs of discarding additional data. Presently, such thresholds are typically selected uniquely for each dataset through a qualitative visual inspection of a variety of outcome

metrics (e.g., QC-RSFC). However, this methodology is largely agnostic to the loss of statistical degrees of freedom relative to the benefits of removal of motion artifact, and thus does not provide a quantitative target that may be used to evaluate the optimality of volume censoring parameters.

In order to address this issue, we sought to develop a benchmark for the data quality improvement resulting from motion denoising in RSFC studies that accounts for the loss of power that is expected to occur as more and more data is censored. Our goal was to optimize the ability of investigators to resolve true sample-level mean RSFC correlations by simultaneously minimizing both a) motion-induced bias, defined as the change in sample mean RSFC correlations, due to targeted volume censoring; and b) the increase in RSFC confidence interval widths resulting from loss of high-motion volumes, runs, and subjects. To this end, we developed a measure of the overall improvement in the quality of RSFC estimates resulting from motion denoising by employing a mean-squared error (MSE) calculation from a bias-variance decomposition (that is, an MSE calculation that includes bias, in addition to variance, in its estimation of total error), divided by the sample size (number of subjects remaining after volume censoring), which we term ΔMSE-RSFC. The results of employing this approach in the HCP500 dataset to determine optimal parameters for LPF-FD and GEV-DV censoring (Φ_F and d_G) when used independently is shown in **Figure 6**. As shown in **Figure 6a**, increasingly aggressive volume censoring (nearly) continuously reduces the motion-induced systematic sample-wide bias in RSFC correlations, with significantly greater effects in analyses without using GSR, using both LPF-FD and GEV-DV volume targeting methods. Additionally, **Figure 6b** shows that both censoring methods produce a reduction in between-subjects variance that exceeds the increase caused by an increase in sampling error until approximately 50% of volumes are

removed, depending on the censoring and analysis methods used (LPF-FD vs. GEV-DV, and with/without GSR). As variance is reduced considerably before any subjects are removed in GEV-DV censoring (**Figure 6c**), this reduction in variance is not due to excluding individuals and a consequent change in sample composition; rather, it is due to the exclusion of volumes that were impacting individual RSFC correlation estimates in relatively high-motion subjects. The resulting ΔMSE-RSFC, calculated as the change in the ratio of the sum of squared bias and variance to the number of remaining subjects (see *Methods*), produces a U-shaped curve when plotted again percent frames removed (**Figure 6d**), and is thus suitable for optimization. The global minimum (i.e., the greatest magnitude reduction in MSE-RSFC) represents the point at which maximal improvement in data quality is achieved in this dataset for each particular method when used in isolation: beyond this point, while further removal of data will result in a reduction in bias, that reduction will be accompanied by a *larger* increase in variance that results in a poorer overall estimate of the true value of RSFC correlations in the sample. The optimal volume censoring parameter and corresponding percent of volumes removed is shown for each method with and without GSR in **Table 1**.

Table 1. Optimal parameters for volume censoring determined by minimizing *ΔMSE-RSFC.*

	No GSR	**With GSR**
LPF-FD Volume Censoring Only		
LPF-FD Φ_F (mm)	0.0318	0.0337
% Frames Removed	46.68	43.32
GEV-DV Volume Censoring Only		
GEV-DV d_G (dimensionless)	1.16	1.36
% Frames Removed	45.95	37.95
Combined LPF-FD and GEV-DV Volume Censoring		
LPF-FD Φ_F (mm)	0.0339	14.336
GEV-DV d_G (dimensionless)	1.39	1.30
Φ_F / d_G ratio (mm)	0.02438	11.0277
Frames Removed (%)	56.15	40.06

Note. GSR = global signal regression. LPF-FD = low-pass filtered framewise displacement. GEV-DV = generalized extreme value derivative of timecourses, root-mean-squared VARiance over voxels.

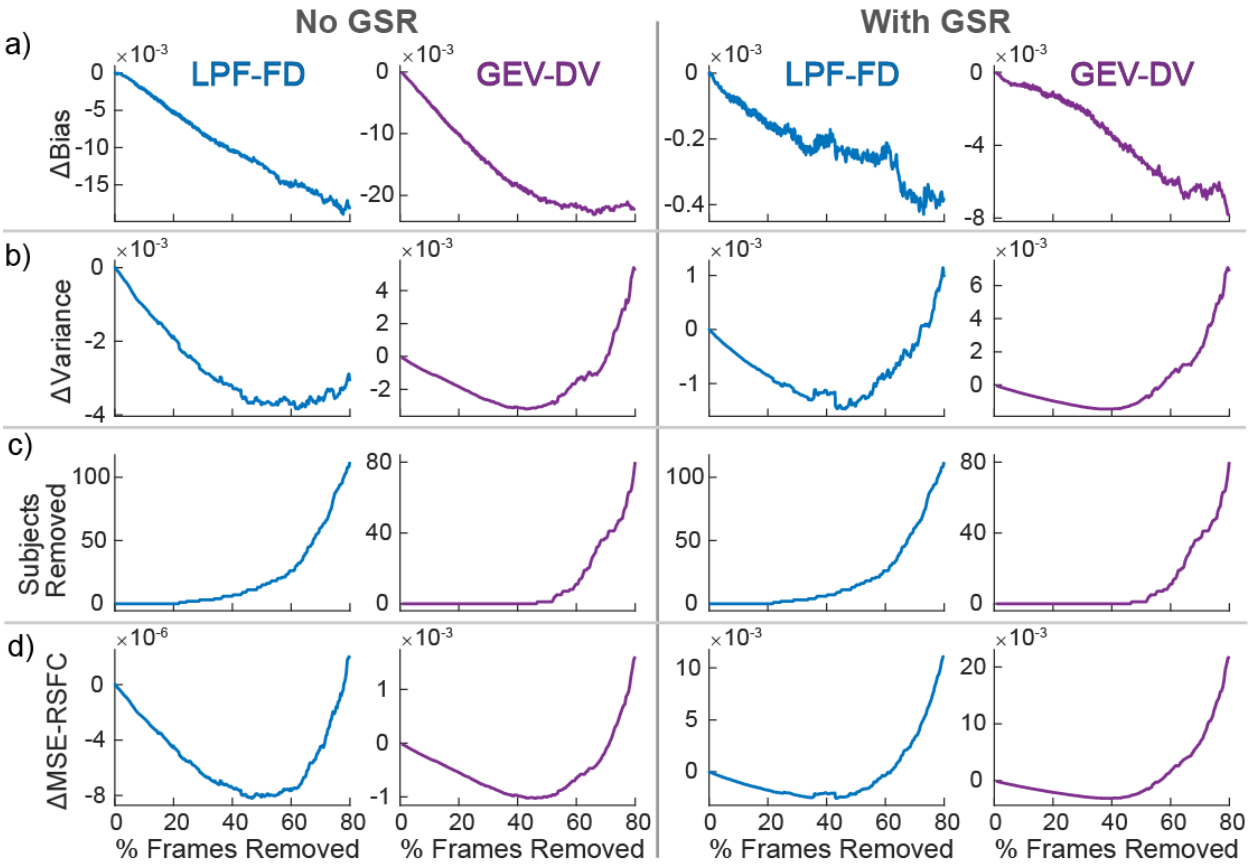

Figure 6. Change in sample statistics due to volume censoring. **a)** Mean change in sample average Z-transformed ROI pair correlations, over and above that due to random removal of an equivalent number of randomly selected volumes within each run. **b)** Mean change in between-subjects variance in sample ROI pair correlation estimates. **c)** Number of subjects removed from the sample due censoring. **d)** Change in estimated mean square error (MSE) divided by the remaining number of subjects, ΔMSE-RSFC, averaged across ROI pairs.

Next, we aimed to determine optimal combined thresholds for LPD-FD and GEV-DV censoring when used together, by determining the location of the global minimum $\Delta MSE\text{-}RSFC$ in the space produced by the free parameters LPF-FD Φ_F and GEV-DV d_G. Our approach was to use a two-parameter global optimization consisting of a particle swarm optimization followed by simulated annealing, each with a pattern search local optimization hybrid function to refine local optima (see *Methods*). Optimal free parameter values for LPF-FD and GEV-DV volume censoring when used simultaneously in the HCP 500 dataset are shown in **Table 1**. These results suggest that a relatively restrictive threshold for LPF-FD is required for analysis without GSR, but that it is optimal to rely primarily on GEV-DV censoring for data employing GSR, which is

consistent with the minimal reductions in bias observed for LPF-FD with GSR in **Figure 6a**. That is, it is likely that the artifactual signals that are indexed by increasing LPF-FD values, but that are not *also* accompanied by outlying values of LPF-DV at the same timepoint, are largely removed using GSR. Consequently, when used in concert with GSR, LPF-FD values can be largely disregarded so long as relatively aggressive GEV-DV censoring is carried out.

Estimation of Optimal Volume Censoring Thresholds for Other Datasets

In order to allow investigators to approximate which thresholds are optimal for a given study protocol (i.e., a given number of runs per subject, and volumes per run) without having to carry out the computationally intensive optimization procedures employed here, we sought to generalize these optimized LPD-FD and GEV-DV thresholds beyond the HCP 500 dataset. Using the known relationship between a) the sampling error of Z-transformed Pearson's correlations calculated within each run, and b) the number of observations (i.e., volumes) within each run and the number of runs acquired per subject, we estimated what the change in between-subjects variance (as shown in **Figure 6b**) would be in a dataset with a different study protocol and an identical distribution of motion and motion artifact across acquired volumes (i.e., volumes and runs are removed in identical proportion to the HCP 500 dataset, with proportional effects on bias and variance; see *Methods*).

Using the protocol-adjusted change in variance we estimated the optimal aggressiveness of volume censoring and the associated LPF-FD and GEV-DV parameters across a range of volumes per run and number of runs per subject. These are shown for each method in isolation in **Figure 7a-b** and **Figure 7c-d**, respectively, as well as for combined LPF-FD + GEV-DV censoring in **Figure 7e-f**. For combined censoring, LPF-FD and GEV-DV were varied along the vector defined by the ratio of LPF-FD Φ_F and GEV-DV d_G determined to be optimal for the

HCP500 dataset in **Table 1**, as demonstrated in **Figure 7g**. Notably, in the absence of GSR, combined LPF-FD + GEV-DV censoring results in improvements to data quality when using a more aggressive combined threshold than those used for LPF-FD or GEV-DV individually, i.e., it is optimal to use both methods together to remove more volumes than when using either

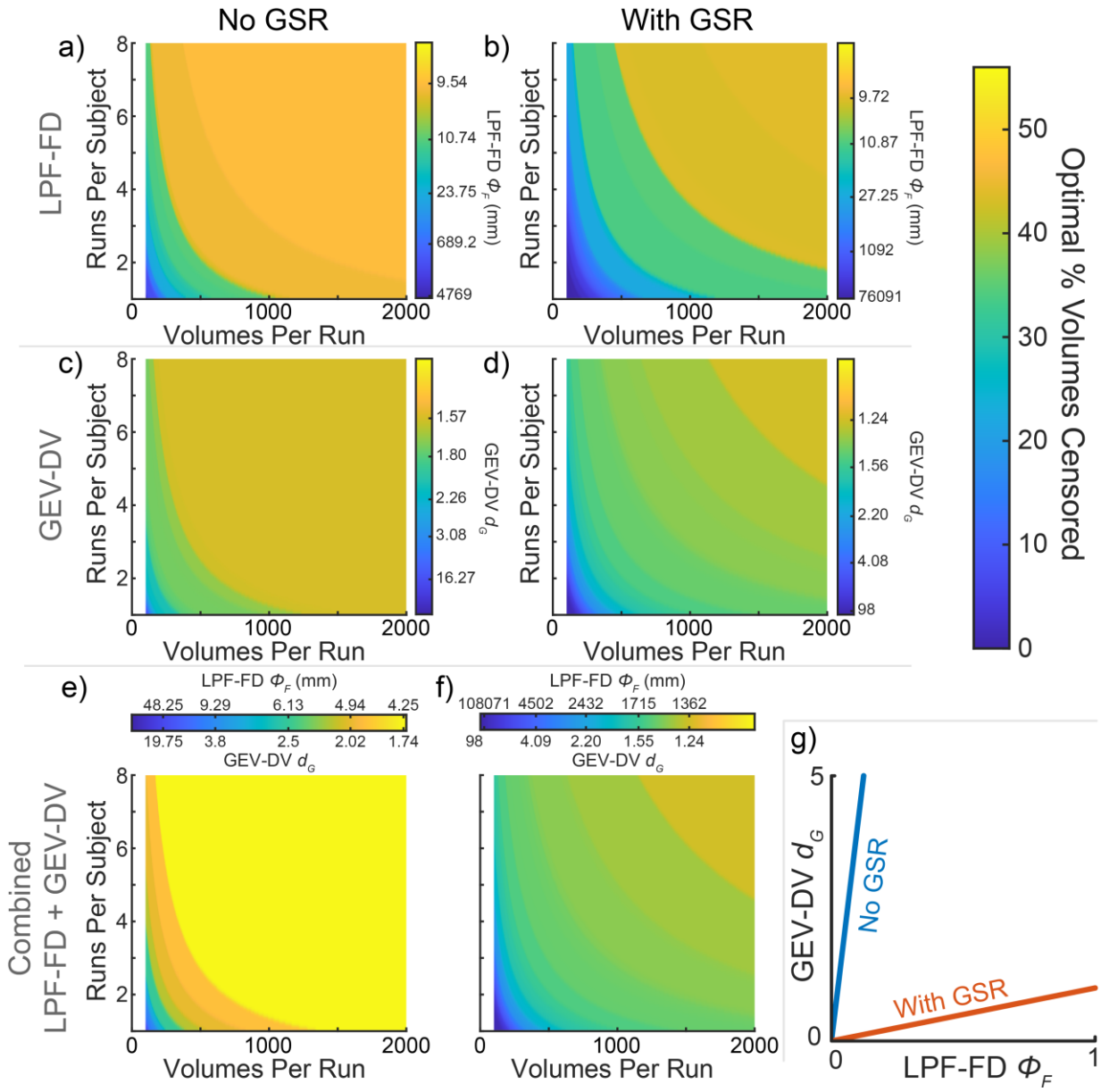

Figure 7. Optimal percent of volumes censored as a function of fMRI acquisition protocol (runs/subject and volumes/run), determined using analytic adjustments to empirically observed *ΔMSE-RSFC* (see *Methods*). Results are shown for LPF-FD **(a,b)**, GEV-DV **(c,d)**, and combined LPF-FD + GEV-DV censoring **(e,f)**, both without GSR **(a,c,e)** and with GSR **(b,d,f)**. **Panel g** demonstrates the optimal ratio of free parameters for combined LPF-FD + GEV-DV censoring from **Table 1** used to produce data in **e-f**. All contour plots **(a-f)** are shown using the same color scale (Optimal % Volumes Censored); each panel displays the corresponding volume censoring parameter values.

method individually. In contrast, optimal data quality when using GSR is produced when using a threshold that relies almost entirely on GEV-DV, while removing fewer volumes than when using LPF-FD alone, in concordance with the results in **Table 1**.

Finally, we aimed to utilize the findings in **Figure 7e-g** in order to provide estimates for optimal volume censoring parameters, for volume censoring using LPF-FD and GEV-DV in tandem, as a function of dataset size (i.e., number of runs per subject and volumes per run) that may be used by investigators for virtually any study protocol. The contribution of within-run variance in Z-transformed RSFC correlations to the between-subject variance at the study level can be calculated as $\frac{1}{N_R \cdot (N_V - 3)}$, where Nv is the number of volumes per run, and N_R is the number of runs per subject that are included in the calculation of RSFC correlations (i.e., after discarding volumes prior to steady state T1 relaxation, or discarding volumes at the beginning and end of a run timeseries following bandpass filtering). Using this relationship, we determined the optimal GEV-DV d_G as a function of this term, for a hypothetical dataset with N_{R_H} runs per subjects and N_{V_H} volumes per run prior to volume censoring. We then fit a monoexponential curve ($A \cdot \exp(b \cdot x) + c$) to this relationship for data without GSR (adjusted R^2 = 0.99960) and with GSR (adjusted R^2 = 0.99999), as shown in **Figure 8a-b** (optimal fit parameters are shown in **Table S1)**. **Table 2** and **Figure 8c-f** show these results for GEV-DV d_G and LPF-FD Φ_F as a function of $N_{R_H} \cdot (N_{V_H} - 3)$, the form of $A \cdot \exp(b/x) + c$, with parameters from the curve fit in **Figure 8a-b**. The functions for optimal LPF-FD thresholds were determined by scaling the magnitude of the monoexponential (parameter A) by the ratio of Φ_L/d_G found to be optimal, separately for analyses with and without GSR (from **Table 1** and **Figure 7g**). **Table 2** presents these results in a straightforward, directly usable form that investigators may use to estimate optimal censoring parameters, simply by plugging in the relevant values for N_{R_H} and N_{V_H}.

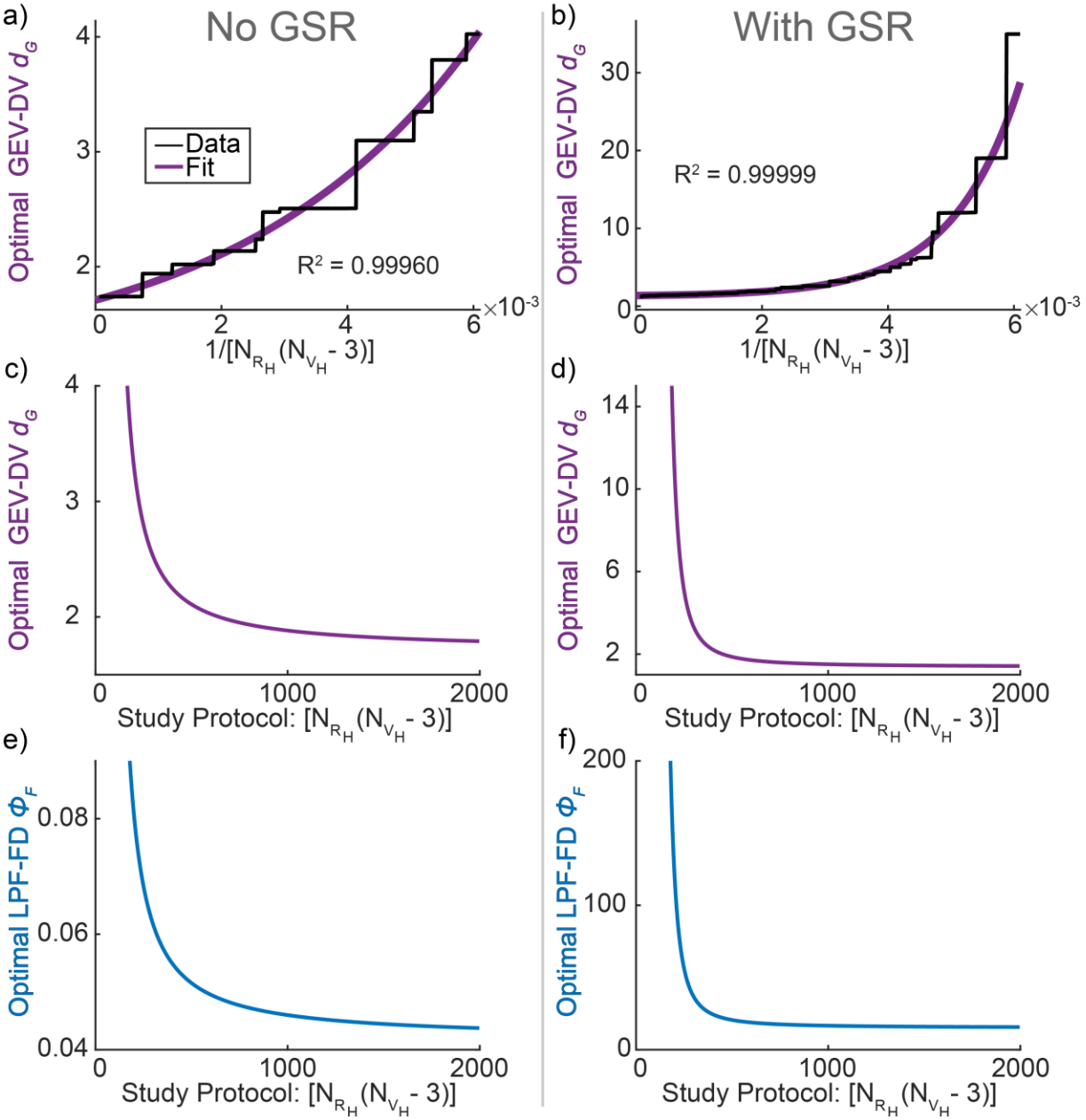

Figure 8. Procedure for estimating optimal censoring parameters as a function of study protocol. **a,b)** Fit of optimal GEV-DV *d*G vs. sampling variance observed at the between-subjects level. Data is fit to an equation of the form: A·exp(b·x) + c, where x = 1/(N_{R_H}·(N_{V_H} - 3)), and A, b, and c are free parameters. Optimal fit parameters are shown in **Table S1**. **c-f)** Relationship between optimal GEV-DV d_G **(c,f)** and LPF-FD Φ_F **(e,f)** and study protocol, i.e., the reciprocal of between-subjects level sampling variance, the product of number of runs (N_{R_H}) per subject and the number of volumes per run minus 3 (N_{V_H} – 3) in the full uncensored dataset. Optimal LPF-FD Φ_F **(e,f)** was obtained by scaling the curve in panels **e-f** by the relationship shown in **Fig. 7g** and **Table 1**.

Table 2. Optimal volume censoring parameters as a function of dataset size.

No GSR	With GSR
GEV-DV d_G (dimensionless)	
$\boldsymbol{d_G} = 0.5591 \cdot \exp\left(\frac{269.6}{N_{R_H} \cdot (N_{V_H} - 3)}\right) + 1.15$	$\boldsymbol{d_G} = 0.0903 \cdot \exp\left(\frac{937.3}{N_{R_H} \cdot (N_{V_H} - 3)}\right) + 1.277$
LPF-FD Φ_F (mm)	
$\boldsymbol{\phi_F} = 0.0136356 \cdot \exp\left(\frac{269.6}{N_{R_H} \cdot (N_{V_H} - 3)}\right) + 1.15$	$\boldsymbol{\phi_F} = 0.995800 \cdot \exp\left(\frac{937.3}{N_R \cdot (N_{V_H} - 3)}\right) + 1.277$

Note. Optimal GEV-DV d_G and LPF-FD Φ_F are shown as a function of study protocol, $N_{R_H} \cdot (N_{V_H} - 3)$, as shown in **Figure 8c-d** and **Figure 8e-f**, respectively. The optimal Φ_F/d_G ratio is demonstrated visually in **Figure 7g**.

Discussion

The work presented here provides several significant advances in preprocessing strategies for reducing the impact of participant motion in RSFC studies employing acquisitions with multiband acceleration and fast repetition times (TRs). First, we proposed two novel methods for evaluating the quality of data in each frame of an RSFC timeseries (LPF-FD and GEV-DV) for volume censoring and demonstrated that these methods outperform existing methods across the board. Next, we developed a novel metric for empirically determining an optimal threshold for both LPF-FD and GEV-DV in isolation, as well as when used in combination, which we term ΔMSE-RSFC. We then analytically extended this metric to provide formulas that estimate optimal thresholds for virtually any dataset, as a function of the number of runs and number of volumes within each run of a dataset. MATLAB code that provides these optimal thresholds, as well as code that calculates LPF-FD and GEV-DV on motion parameter estimates and whole-brain RSFC timeseries (i.e., NIFTI or CIFTI format images of the RSFC data for a run), respectively, is publicly available (see *Code Availability*).

In addition, this work also provides two key advances for investigators engaged in developing and evaluating preprocessing strategies for motion correction in RSFC. First, we show that some widely used approaches to evaluate the performance of preprocessing strategies for mitigating motion effects on RSFC estimates rest on the flawed assumption that *true*,

'artifact-free', RSFC correlations are unrelated to motion QC measures, such as the median FD value for a participant. Specifically, we demonstrated that when evaluating the number of significant differences in RSFC correlations between high- and low-motion subjects (which has commonly been used to evaluate how much residual motion artifact remains in a dataset after preprocessing), removing participants who have a family history of psychiatric or neurological disorder (FH+), or who tested positive for substance use or high blood alcohol level during the study (SU+), results in a significantly larger reduction in the number of significant differences between high- and low-motion subjects than is observed when removing FH- and SU- participants who exhibit an equivalent amount of in-scanner motion. Thus, we conclude that many of the observed significant differences in RSFC correlations between high- and low-motion subjects in the full dataset actually reflect *true* differences in RSFC between FH/SU+ and FH/SU- individuals, rather than reflecting the presence of motion artifact. This finding implies that the correct interpretation of a change in the number of observed significant differences in RSFC between high- and low-motion participants due to volume censoring is uncertain, and thus problematic as a quantitative measure for evaluating performance, as higher quality motion-correction pipelines may reveal true neural differences in RSFC while simultaneously removing those that are spurious.

In order to address this issue, we also developed a novel metric for evaluating motion denoising pipelines: MAC-RSFC. This metric simply measures the mean absolute change in Pearson correlation in an average RSFC ROI-pair across participants in a dataset as a result of motion-targeted censoring, compared to the same change observed when randomly censoring an equivalent number of volumes. Thus, to the extent that this measure increases in one pipeline compared to another for any given number of volumes censored, it can be inferred that a larger

impact on the data has been had specifically as a result of censoring motion-impacted volumes in the timeseries.

Caveats and Further Considerations

We wish to highlight that the novel methods presented here were specifically designed for multiband data, and we would strongly advise against generalizing these findings to any single band, slower TR, dataset, without further evaluation. However, although these methods were developed specifically on the publicly available HCP 500 subjects data release, we do expect that they will generalize broadly to other multiband datasets, at least provided that they are of similar TR. Substantive differences in TR from the 720 ms used in the HCP study may produce changes in the magnitude of measured LPF-FD and LPF-DV values that cause the thresholds reported here to no longer be appropriate. However, provided the TR of a dataset is reasonably close to 720 ms, we expect that the thresholds we report here should be approximately correct; certainly, we anticipate that they would outperform any other volume censoring procedure currently available in the literature.

In addition, it is critical to note that investigators should not assume that applying the formula we provide for optimal censoring thresholds (see **Table 2**) necessarily results in adequate removal of motion artifact from a dataset. Rather, because our method attempts to balance measurement error (which depends in part on the quantity of data remaining after censoring) and motion-induced bias (which depends on the stringency of censoring thresholds), datasets with very brief RSFC acquisitions, and thus relatively few timepoints even prior to censoring, will require very lenient thresholds because variance increases so rapidly as data is removed. Rather, **Figure 7** suggests that a *minimum* of approximately 1,000 volumes (12 minutes of data collection at the 720 ms TR used here) is required before recommended

thresholds begin to level off, consistent with this being the point at which more aggressive censoring begins to have smaller marginal returns in reducing motion-induced bias remaining in the data. Below this quantity of data, the loss of power associated with more aggressive censoring thresholds prohibits the use of sufficiently aggressive LPF-FD and GEV-DV cutoffs, and consequently leaves substantial motion artifact in the dataset.

Further, as shown in **Figure S3** and discussed in *Supplementary Results*, the distribution of motion artifact systematically varies across runs within each session, and over time within each run, with the highest-quality data acquired near the start of each run, and during the first run of each session. Notably, a substantial "rebound" towards higher quality data occurs at the beginning of the second run, and thus it is advisable to aim to employ a larger number of shorter runs rather than a smaller number of long runs. Specifically, **Figure S3** strongly suggests that employing twice as many runs at half the length employed in the HCP study would have resulted in substantially less participant motion than was observed.

Another consideration is that we did not separately consider spike regression methods, because although they differ from censoring in implementation (direct removal of high-motion volumes from the timeseries in the latter, compared to inclusion of a nuisance regressor for each effected volume in the former), the effect of censoring versus spike regression on RSFC correlations and degrees of freedom is identical. Thus, observed differences between these two approaches[2,3] are a result of differences in how high-motion volumes are identified, rather than the removal method per se. Because the methods have been shown to perform similarly[2,3], we felt that comparing our approach to only one of these two methods would enhance clarity and simplify presentation of our results, with no substantive impact on our conclusions.

Methods

Resting State Functional Connectivity Dataset and Pre- and Post-Processing

We employed minimally preprocessed data[34] from the HCP 500 Subjects data release, comprising 501 subjects, each with 4 runs (or, in rare cases, 2 runs) of resting-state fMRI data of 14 minutes 24 sseconds (1200 volumes), collected over two sessions, and described in detail elsewhere[13,24,35]. Data were acquired using the HCP 3T Siemens "Connectome Skyra" scanner, using 2 mm^3 isotropic voxel size, 720 ms TR, and multiband acceleration factor of 8[13,35]. All subsequent analyses were conducted using custom MATLAB functions. RSFC processing was kept generally similar to previously published methods[5], with most changes designed to reduce computational and disk utilization demands.

A set of 10 mm diameter spheres were drawn around 264 center coordinates reported elsewhere[25], and fMRI timeseries were averaged across all voxels in each sphere to generate 264 timeseries for each run. Global signal (GS) was calculated as the mean signal in all voxels in the brain, determined for each subject (from the brainmask_fs.2.nii.gz files included with HCP subject data). Nuisance signals from white matter (WM) and cerebrospinal fluid (CSF) were calculated as the average signal in all compartment voxels remaining after an iterative erosion procedure, in which masks were eroded up to four times as long as some voxels remained in the mask following erosion[5].

The first 10 volumes of each run timeseries were discarded, as inspection of mean timeseries values averaged across all participants showed elevated signal in these volumes (data not shown), consistent with an extended tissue relaxation effect remaining in the data. Next, all timeseries were mode 1000 normalized (multiplied by 1000 and divided by the modal value of all in-brain voxels), demeaned, and detrended. All timeseries were band-pass filtered between

0.009 and 0.08 Hz using a second-order zero-phase Butterworth filter, and the first and last 30 volumes of each timeseries were discarded (as these volumes showed systematic deviations from expected when averaged across the entire dataset, due to filter effects at the edges of the timeseries).

Resting-State Functional Connectivity Analysis Methods

RSFC correlations were calculated using partial correlations, by comparing each pairwise ROI while partialing out the effect of various nuisance parameters. Nuisance parameters included band-pass filtered motion parameters (MPs) and their squares, the derivatives of the band-pass filtered MPs and the squares of those derivatives, the WM signal and its derivative, and the CSF signal and its derivative. All derivatives were calculated by backwards difference. Analyses both with and without GSR (i.e., with and without including mean signal in all brain voxels, and its first derivative) were conducted and are presented side by side throughout. All RSFC correlations were Fisher's *r*-to-*Z* transformed immediately after their calculation, prior to their use in any further computations or analysis, and are reported in this form throughout this manuscript. In all cases, (Z-tranformed) RSFC correlations for each subject were calculated separately for each available fMRI run (4 for most participants, although some HCP participants have fewer) and averaged together across runs.

Standard Volume Censoring Methods

For standard (FD and DV) volume censoring, high-motion volumes were identified and targeted for removal by volume censoring (scrubbing) when their FD exceeded a threshold FD value, or when their DV value exceeded a threshold DV value. FD was calculated as the

estimated motion of a cortical voxel from one frame to the next, based on translation and the rotation of a point on the circumference of a 50 mm radius sphere[4]; DV was calculated as the root-mean-square (over voxels) of the first derivative (by backwards differences) of the timeseries across all brain voxels[4,23]. Calculation of standard DV values was carried out following volume smoothing with a 4 mm FWHM gaussian kernel in SPM 12, in order to produce values closer to those reported for prior datasets[4,5], as unsmoothed data produces much higher values of DV. Prior to band-pass filtering of timeseries, linear interpolation was used to replace censored time series data before discarding these time points from analysis. Runs for which less than 2 minutes of uncensored data (167 volumes) remained after censoring were excluded from analysis. Study participants were retained for analysis provided they had at least 1 run remaining after censoring.

LPF-Based Volume Censoring (LPF-FD and LPF-DV) and Adaptive LPF-DV Censoring (GEV-DV)

LPF-FD was calculated for each frame by calculating FD as above, but on a set of MPs that were first low-pass filtered at 0.2 Hz with a second-order Butterworth filter. LPF-DV was calculated by applying the same low-pass filter to voxel timeseries data prior to calculation of DV (as above). Similar to standard FD and DV volume censoring, the aggressiveness of LPF-FD and LPF-DV volume censoring was set by selecting threshold LPD-FD and LPF-DV values, respectively, denoted Φ_F and Φ_D.

Adaptive LPF-DV censoring thresholds were set by maximum likelihood fitting of a generalized extreme value (GEV) distribution[36] to the LPF-DV values within each run, and setting the threshold LPF-DV value Φ_D separately within each run such that the cumulative

density function (CDF) at Φ_D is equal to $1 - \frac{k_G+0.3}{d_G}$ (i.e., the area under the curve of the GEV to the right of the cutoff is equal to $\frac{k_G+0.3}{d_G}$), where k_G is the shape parameter obtained from the GEV fit, and d_G is a free parameter. The shape parameter k_G is greater in runs containing more extreme DV values (i.e., a thicker right tail), causing a greater proportion of the data to be excluded when more data has high LPF-DV values relative to the central tendency for that run. The free parameter d_G allows investigators to set the overall aggressiveness of the cutoff across the dataset, which is used in place of a fixed LPF-DV threshold. In cases where LPF-FD and GEV-DV volume censoring are used together, volumes with an LPF-FD value exceeding the threshold set for the study (Φ_F), or an LPF-DV value exceeding the threshold set for the run (Φ_D), were censored (i.e., a logical OR operation on the censoring vectors produced by each method).

Comparison of High- to Low-Motion Participants After Censoring

We also sought to compare the number of significant differences in RSFC correlations between high- and low-motion participants, following prior work[3,5,7,9], by splitting the sample into terciles based on median FD for each participant. However, as described in the results, we observed an alarmingly high number of significant differences between high- and low-motion participants following motion censoring. This led us to hypothesize that the assumption that this approach rests on (that there are no true RSFC differences between high- and low-motion participants, such that any such differences can be ascribed to the effects of motion) might be false. Consequently, we used the HCP Tier 2 Restricted Data release to identify and remove participants with a family history of any psychiatric or neurological condition (e.g. schizophrenia, Parkinson's disease, etc.), or who engaged in drug use or had elevated blood-

alcohol content (BAC) in the course of the study, which resulted in exclusion of 154 and 32 participants, respectively.

In order to determine whether removal of these participants resulted in a greater-than-expected reduction in the number of differences in RSFC correlations, we also generated 1000 Monte Carlo resamplings of the dataset that excluded the same number of participants who did not have a family history of psychiatric or neurological disorder, or positive toxicology or elevated BAC, but who still had approximately the same mean motion as the originally excluded participants. This was done by restricting the set of participants from which resampling could occur to those with a median FD value across all RSFC runs that was either higher or lower than the "true" excluded sample mean, depending on whether the current sample mean was higher or lower than the mean of the true sample. That is, for each Monte Carlo iteration, a new set of 186 participants to be removed was randomly selected, one participant at a time. The first participant to be removed was randomly selected from the entire set of participants who lack a family history of psychiatric or neurological disorder, positive toxicology, and elevated BAC. If this participant's median FD value was greater than the mean of the true 186 excluded participants, then the second participant was randomly selected from only those participants with a median FD value lower than the mean of the true 186 participants; conversely, if the participant's median FD value was lower than the mean of the true sample, then the second participant had to have a median value that was higher than the mean. The third participant was then selected from those participants with higher or lower median FDs than the true sample mean, based on whether the mean of the first two participants was higher or lower than the true sample mean. This process was repeated iteratively until a full sample of 186 participants had been constructed, for each of the 1000 Monte Carlo resamplings.

Significant differences between high- and low-motion participants (the upper and lower tercile of the median FD for each participant across all volumes in all runs, respectively) were then determined for each ROI pair at a range of uncorrected p-value thresholds (ranging from $P < 0.0001$ to $P < 0.1$, in increments of 0.00001). This was done for the full dataset, for a dataset excluding all family history or substance use positive participants, and for each of the 1000 Monte Carlo samples.

Evaluation of Volume Censoring Performance

Changes to RSFC correlations as a result of targeted volume censoring were evaluated for each method as compared to randomly censoring an equivalent number of volumes within each run ("random censoring"). Contiguous clusters of "bad" volumes (i.e., volumes flagged for removal) were randomly permuted, in order to maintain the size and number of censored regions in each run, and averaged over 10 randomizations[4]**.**

Figure 5 shows a comparison of volume censoring methods across a range of parameter values for each method. The range of parameters evaluated was as follows. For FD and LPF-FD (all units in mm): 1001 points from 0 though 0.1 in steps of 10^{-4}, 160 points from 0.1025 through 0.5 in steps of 2.5×10^{-3}, 50 points from 0.51 through 1.0 in steps of 10^{-2}, 90 points from 11 through 100 in steps of 0.1, and infinity (corresponding to no censoring). For DV and LPF-DV (all units in tenths of % signal change per frame): 1001 points from 0 through 10 in steps of 0.01, 160 points from 10.25 through 50 in steps of 0.25, 50 points from 51 through 100 in steps of 1, 90 points from 110 through 1,000 in steps of 10, and infinity (corresponding to no censoring). For GEV-DV d_G (all units dimensionless): 1,000 points from 0.005 through 5 in steps of 0.005,

50 points from 5.1 through 10 in steps of 0.1, 90 points from 11 through 100 in steps of 1, and infinity (no censoring).

The mean absolute change in resting state functional connectivity (MAC-RSFC) was determined by first calculating the mean within-subject RSFC correlations in all ROI pairs after targeted volume censoring, and subtracting the mean RSFC correlations resulting from random censoring (averaged over 10 random censoring vectors), to obtain $\Delta Z_{i,k}$ for the *i*th subject and *k*th ROI pair. This is equivalent to subtracting the change in RSFC correlations produced by targeted volume censoring from that due to random censoring.

That is, for subject *i*, run *j*, ROI pair *k*, and random permutation vector *p*

$$\Delta Z_{i,k} = \sum_{j=1}^{N_{R_i}} \frac{Z_{i,j,k} - \frac{\sum_{p=1}^{N_{RAND}} Z_{RAND\,i,j,k,p}}{N_{RAND}}}{N_{R_i}}, \quad (1)$$

where $Z_{i,j,k}$ is the RSFC correlation for subject *i*, run *j*, and ROI pair *k*, after targeted volume censoring, $Z_{RAND\,i,j,k,p}$ is the RSFC correlation for subject *i*, run *j*, and ROI pair *k* after volume censoring using random permutation vector *p*, N_{R_i} is number of runs that exist after censoring for subject *i*, and the number of random permutations used per run is N_{RAND} (here, 10). MAC-RSFC is calculated for the full sample as

$$\text{MAC-RSFC} = \sum_{k=1}^{N_{PAIR}} \frac{\sum_{i=1}^{N_S} \frac{|\Delta Z_{i,k}|}{N_S}}{N_{PAIR}}, \quad (2)$$

in which N_S is the number of subjects in the sample after volume censoring and N_{PAIR} is the number of ROI pairs used for analysis (here, $N_{PAIR} = \binom{264}{2} = 34{,}716$). 95% confidence intervals for MAC-RSFC were approximated for each parameter value using the bias-corrected and accelerated (BCa) bootstrap with 10,000 bootstrap samples[37]. Note that MAC-RSFC has a

nonzero null expectation, making it unsuitable for evaluating whether a given parameter value for a given method has a significant non-zero effect. However, comparisons between parameter values within each method, or between two or more methods, do have a zero null expectation. That is, in the absence of any true between-method (or between-parameter) differences in the impact of targeting volumes for removal on RSFC estimates, MAC-RSFC would be expected to be identical across methods. Note further that, while *ΔMSE-RSFC* (see below) decomposes the effects of volume censoring on sample RSFC estimates into bias and variance components, MAC-RSFC measures both in aggregate.

Delta-Mean-Squared-Error Based Methods (ΔMSE-RSFC)

We developed *ΔMSE-RSFC* as a quantitative benchmark for volume censoring performance that could be used as an optimization target to determine optimal free parameter values for each volume censoring method (i.e., optimal LPF-FD Φ_F and GEV-DV d_G). As explained in detail in *Results*, we evaluated the mean squared error (MSE) in RSFC correlations for each pair of ROIs, divided by the number of subjects remaining after censoring (N_S), over the range of parameter values noted above for LPF-FD and GEV-DV. *ΔMSE-RSFC* was calculated in each ROI pair *k* and averaged over all ROI pairs, i.e.,

$$\Delta MSE\text{-}RSFC = \frac{\sum_{k=1}^{N_{PAIR}} \Delta \text{MSE-RSFC}_k}{N_{PAIR}}, \quad (3)$$

where N_{PAIR} is the number of ROI pairs (34,716 as above), and

$$\Delta MSE\text{-}RSFC_k = \Delta\left(\frac{MSE(\hat{\theta}_k)}{N_S}\right) = \frac{MSE(\hat{\theta}_{C_k})}{N_S} - \frac{MSE(\hat{\theta}_{U_k})}{N_{S_U}}, \quad (4)$$

where $\hat{\theta}_{U_k}$ is the estimated RSFC correlation in ROI pair *k* from uncensored data, $\hat{\theta}_{C_k}$ is the same estimator for volume censored data, and N_{S_U} is the number of subjects in the dataset in the

uncensored dataset. $MSE(\hat{\theta}_k)$ is the mean squared error in the RSFC correlation estimate for ROI pair *k*, accounting for both the variance (observed across subjects) and bias produced by motion artifact

$$MSE(\hat{\theta}_k) = E[(\hat{\theta}_k - \theta_k)^2] = \sigma^2(\hat{\theta}_k) + Bias(\hat{\theta}_k, \theta_k)^2, \quad (5)$$

in which θ_k is the true population mean RSFC correlation for a single ROI pair *k* and $\hat{\theta}_k$ is its estimator, $\sigma^2(\hat{\theta}_k)$ is the observed between-subjects variance in $\hat{\theta}_k$, and $Bias(\hat{\theta}_k, \theta_k)$ is the bias in $\hat{\theta}_k$, here defined as the total motion-induced bias that is removeable by volume censoring (see elsewhere[38] for further details on the bias-variance decomposition of *MSE*). Thus,

$$\Delta MSE\text{-}RSFC_k = \frac{\sigma^2(\hat{\theta}_{C\,k}) + Bias(\hat{\theta}_{C\,k}, \theta_k)^2}{N_S} - \frac{\sigma^2(\hat{\theta}_{U\,k}) + Bias(\hat{\theta}_{U\,k}, \theta_k)^2}{N_{S_U}}. \quad (6)$$

As it is not possible to measure $Bias(\hat{\theta}_k, \theta_k)$ directly, we instead estimated the change in bias due to volume censoring, $\Delta Bias(\hat{\theta}_k, \theta_k)$, in each ROI pair by measuring the mean sample-wide magnitude of the change in its RSFC correlation due to volume censoring and taking the additive inverse, i.e.,

$$\Delta Bias(\hat{\theta}_k, \theta_k) = -\left|\sum_{i=1}^{N_S} \frac{\Delta Z_{i,k}}{N_S}\right|, \quad (7)$$

where $\Delta Z_{i,k}$ is defined in Equation 1. The change in squared bias can then be estimated as

$$\Delta\left[Bias(\hat{\theta}_k, \theta_k)^2\right] = \left|\left[Bias(\hat{\theta}_{U\,k}, \theta_k) + \Delta Bias(\hat{\theta}_k, \theta_k)\right]^2 - Bias(\hat{\theta}_{U\,k}, \theta_k)^2\right|, \quad (8)$$

and thus

$$Bias(\hat{\theta}_{C\,k}, \theta_k)^2 = Bias(\hat{\theta}_{U\,k}, \theta_k)^2 + \Delta\left[Bias(\hat{\theta}_k, \theta_k)^2\right]. \quad (9)$$

As noted above, $Bias(\hat{\theta}_{U\,k}, \theta_k)$ is the total sample-wide motion-induced bias in RSFC correlations that is removeable by volume censoring, and is necessarily not known. However, we

estimate this value for each ROI pair by measuring the additive inverse of the change in mean RSFC correlations due to volume censoring, $\Delta Bias(\hat{\theta}_k, \theta_k)$, as a function of percent of frames removed (as in the top row of **Figure 6**), resampling to a resolution of 0.01% using linear interpolation, estimating the slope using robust regression (bisquare weight function, tuning constant of 4.685) and an intercept of 0, and extrapolating to estimate the value at 100% frames removed. This method was used in order to provide a reasonable estimate of total bias that is robust to the instability in RSFC correlation estimates as the number of frames removed begins to approach 100% (notably, this instability also necessitates the absolute value employed on the right side of Equation 8, so that this instability near 100% frame removal is not treated as the maximal removal of bias). As a demonstration, this method is shown for the average RSFC correlation across all ROI pairs, using LPF-FD based censoring, without GSR, in **Figure S4.**

Determination of Optimal Volume Censoring Parameters Using Multi-Parameter Global Optimization

Optimal parameters for combined LPF-FD and GEV-DV volume censoring were determined by using an optimization procedure to minimize *ΔMSE-RSFC*. Optimization was carried out using a particle swarm global optimization algorithm (particleswarm in MATLAB[39-41]), followed by simulated annealing (simulannealbnd in MATLAB[42-44]), each followed by a pattern search local optimization algorithm (patternsearch in MATLAB[45-50] with default settings) to refine results. Particle swarm optimization was carried out using 44 particles, 40 of which were randomly generated upon initialization in the interval $[10^{-1}, 10^{4}]$, with a $\log_{10}$-uniform distribution, for both LPF-FD Φ_F and GEV-DV d_G. The 4 remaining particles were placed along the edges of this distribution, i.e. $(10^{-1}, 10^{-1})$, $(10^{-1}, 10^{4})$, $(10^{4}, 10^{-1})$, and $(10^{4}, 10^{4})$. Parameters for

the particle swarm optimization were 1000 stall iterations and an initial neighborhood size of 1. Simulated annealing was carried out using the optimal points determined by particleswarm (after refinement of local minima using patternsearch) as starting values, with an initial temperature of 1000, reannealing interval of 100, and function tolerance of 10^{-6}, with steps with length temperature and direction chosen uniformly at random (i.e., 'annealingfast').

Estimation of Optimal Censoring Thresholds for Other Datasets

In order to determine optimal volume censoring parameters appropriate for a variety of acquisition protocols (i.e., protocols that differ from the HCP datasets in number of runs per subject and number of volumes per run), we developed an approach to approximate the *ΔMSE-RSFC* that would be expected over a range of protocols as a function of volume censoring parameters (LPF-FD and GEV-DV).

Our approach was to first consider a variance decomposition of a given RSFC correlation for ROI pair *k*

$$\sigma^2(\hat{\theta}_k) = \sigma_S^2(\hat{\theta}_k) + \frac{\sigma_R^2(\hat{\theta}_k) + \sigma_Z^2(\hat{\theta}_k)}{\tilde{N}_R} + \sigma_A^2(\hat{\theta}_k), \quad (10)$$

where $\tilde{N}_R$ is the harmonic mean of the number of runs across all subjects in the study, $\sigma^2(\hat{\theta}_k)$ is the observed variance across study participants for ROI pair *k*, $\sigma_S^2(\hat{\theta}_k)$ is the true between-subjects variance in the RSFC correlation that exists between individual participants, $\sigma_R^2(\hat{\theta}_k)$ is the true between-runs variance in RSFC correlation *k*, $\sigma_Z^2(\hat{\theta}_k)$ is the within-runs variance in Z-transformed Pearson's product-moment correlations due to sampling error, and σ_A^2 is the total contribution of motion artifact (and potentially other sources of unexplained error) to $\sigma^2(\hat{\theta}_k)$. For simplicity, we assume in these analyses that RSFC is a stationary process, and thus variance

in observed correlations between runs is solely due to $\sigma_Z^2(\hat{\theta}_k)$, i.e., there is no true between-runs variance, and so we get:

$$\sigma^2(\hat{\theta}_k) = \sigma_S^2(\hat{\theta}_k) + \frac{\sigma_Z^2(\hat{\theta}_k)}{\widetilde{N}_R} + \sigma_A^2(\hat{\theta}_k). \quad (11)$$

Further, we can use the known variance of a Z-transformed Pearson correlation (which depends only on the number of observations), such that for subject *i* and run *j*

$$\sigma_Z^2(\hat{\theta}_{i,j,k}) = \frac{1}{N_{V\,i,j}-3}, \quad (12)$$

where $N_{V\,i,j}$ is the number of volumes remaining in the analysis after volume censoring for subject *i* and run *j*. It follows that we can write the variance at the study (between-subjects) level due to $\sigma_Z^2(\hat{\theta}_k)$, averaged across all runs and all subjects, here defined as $\sigma_{Z,STUDY}^2(\hat{\theta}_k)$, as

$$\sigma_{Z,STUDY}^2(\hat{\theta}_k) = \sum_{r=1}^{N} \frac{\sum_{i=1}^{N_{R_i}} \frac{1}{(N_{R_i})(N_{V\,i,j}-3)}}{N_S}. \quad (13)$$

Equation 13 can also be written as a product of harmonic means for the whole dataset,

$$\sigma_{Z,STUDY}^2(\hat{\theta}_k) = \frac{N_S}{\sum_{i=1}^{N_S}\frac{1}{(N_{R,i})}} \cdot \frac{N_S}{\sum_{i=1}^{N_S}\frac{\sum_{j=1}^{N_{R_i}}\left(\frac{1}{N_{V\,i,j}-3}\right)}{N_{R_i}}} = \frac{1}{\widetilde{N}_R(\widetilde{N_V-3})} = \frac{1}{\widetilde{N}_R \cdot \tilde{\nu}}, \quad (14)$$

where $\widetilde{N}_R$ is the across-subjects harmonic mean of N_{R_i}, and $\tilde{\nu}$ is the across-*subjects* harmonic mean of the across-*runs* harmonic mean of $(N_{V\,i,j}-3)$. Further, we can define χ and γ as the proportions of runs and volumes remaining after censoring, respectively, i.e., $\chi = \frac{\widetilde{N}_R}{\widetilde{N}_{R_U}}$, and $\gamma = \frac{\tilde{\nu}}{\tilde{\nu}_U}$, where $\widetilde{N}_{R_U}$ is the harmonic mean of the number of runs across subject before censoring and $\tilde{\nu}_U$ is the harmonic mean of $(N_{V\,i,j}-3)$ across runs and subjects before censoring, such that,

$$\sigma^2_{Z,STUDY}(\hat{\theta}_k) = \frac{1}{(\chi \cdot \tilde{N}_{R_U})(\gamma \cdot \tilde{v}_U)}. \quad (15)$$

As a consequence of **Equation 11**, we can estimate the increase in between-subjects variance that occurs due to removal of volumes and runs as a result of volume censoring as

$$\Delta\sigma^2_{Z,STUDY}(\hat{\theta}_k) = \sigma^2_{Z_C}(\hat{\theta}_k) - \sigma^2_{Z_U}(\hat{\theta}_k), \quad (16)$$

where $\sigma^2_{Z,U}(\hat{\theta}_k)$ and $\sigma^2_{Z,C}(\hat{\theta}_k)$ are calculated for the study per **Equation 15** for the HCP500 dataset before and after volume censoring, respectively. Thus,

$$\Delta\sigma^2_{Z,STUDY}(\hat{\theta}_k) = \frac{(1-\chi)\cdot\tilde{N}_{R_U} + (1-\gamma)\cdot\tilde{v}_U}{(\chi\cdot\tilde{N}_{R_U} + \gamma\cdot\widetilde{v_U})(\tilde{N}_{R_U} + \tilde{v}_U)}. \quad (17)$$

Our approach is to approximate what $\Delta\sigma^2_Z(\hat{\theta}_k)$ would be in a dataset of a different size, i.e., how this term would change for a different $\tilde{N}_{R_U}$ and $\tilde{v}_U$, and use this estimate to adjust our observed estimate of $\sigma^2(\hat{\theta}_k)$ from the actual HCP500 dataset. First, we assumed that in a hypothetical dataset, motion artifact would be distributed identically across frames, runs, and subjects, with equivalent magnitude, as in the HCP500 dataset. Therefore, an equivalent proportion of this hypothetical dataset (i.e., an equal percentage of subjects, runs, and volumes) would be removed for any given volume censoring parameter as was actually removed in the HCP500 dataset. We can estimate the study protocol-adjusted $\Delta\sigma^2_Z(\hat{\theta}_k)$ for any given set of volume censoring parameters for a hypothetical dataset with uniformly N_{R_H} runs per subject and N_{V_H} volumes per run, and $v_H = (N_{V_H} - 3)$,

$$\Delta\sigma^2_{Z,STUDY_H}(\hat{\theta}_r) = \frac{(1-\chi)\cdot N_{R_H} + (1-\gamma)\cdot v_H}{(\chi\cdot N_{R_H} + \gamma\cdot v_H)(N_{R_H} + v_H)}. \quad (18)$$

Consequently, for all estimates of the change in between-subjects variance due to volume censoring, $\Delta\sigma^2(\hat{\theta}_k)$ we can additionally determine the change in variance that should be

observed in a hypothetical dataset, denoted $\Delta\sigma_H^2(\hat{\theta}_k)$, by subtracting the estimated increase in variance due to censoring observed at the between-subjects level in the HCP500 dataset, as estimated in **Equation 17**, and adding that which would have theoretically been observed in a hypothetical dataset as estimated in **Equation 18**

$$\Delta\sigma_H^2(\hat{\theta}_k) = \Delta\sigma^2(\hat{\theta}_k) + \Delta\sigma^2_{Z,STUDY_H}(\hat{\theta}_k) - \Delta\sigma_Z^2(\hat{\theta}_k), \tag{19}$$

and thus,

$$\sigma_H^2(\hat{\theta}_{C_k}) = \sigma^2(\hat{\theta}_{U_k}) + \Delta\sigma_H^2(\hat{\theta}_k). \tag{20}$$

This can then be used to obtain an estimate of *ΔMSE-RSFC* for this hypothetical dataset by replacing $\sigma^2(\hat{\theta}_{C_k})$ terms in **Equation 6** with $\sigma_H^2(\hat{\theta}_{C_k})$ and, assuming an equal proportion of subjects removed due to volume censoring in this hypothetical dataset as in the HCP500 dataset,

$$\Delta MSE\text{-}RSFC_{H_k} = \frac{\sigma_H^2(\hat{\theta}_{C_k}) + Bias(\hat{\theta}_{C_k}, \theta_k)^2}{\zeta N_{S_H}} - \frac{\sigma^2(\hat{\theta}_{U_k}) + Bias(\hat{\theta}_{U_k}, \theta_k)^2}{N_{S_H}}, \tag{21}$$

in which N_{S_H} is the number subjects in the hypothetical dataset, ζ is the proportion of subjects remaining in the dataset after volume censoring, i.e., $\zeta = \frac{N_S}{N_{S_U}}$, where N_{S_U} is the number of subjects in the full, uncensored, HCP500 dataset. Finally, we can calculate the average $\Delta MSE\text{-}RSFC_{H_k}$ across all ROI pairs using **Equation 1**,

$$\Delta\text{MSE-RSFC}_H = \sum_{k=1}^{N_{PAIR}} \frac{\Delta MSE\text{-}RSFC_{H_k}}{N_{PAIR}}. \tag{22}$$

Determination of Optimal Volume Censoring Parameters for Variable Acquisition Protocols

Using **Equations 18-22**, we varied the initial number of frames and runs in a hypothetical dataset where N_{R_H} was varied in the range [1,8] in steps of 0.05, and N_{V_H} was varied in the range [100, 2000], in steps of 1 from 100 through 500, and in steps of 5 from 505 through 2000.

ΔMSE-RSFC was recalculated for each combination of N_{R_H} and N_{V_H}, for volume censoring parameters using the optimal ratio of LPF-FD Φ_F to GEV-DV d_G when used in tandem as shown in **Table 1**, i.e., along the vectors shown in **Figure 7g**, separately for analyses with GSR and without GSR. A total of 3,582 sets of volume censoring parameters were sampled, using the union of the set of parameters used to sample LPF-FD and GEV-DV independently when calculating MAC-RSFC and *ΔMSE-RSFC* for the HCP500 dataset, and the aforementioned optimal FD Φ_F/d_G ratio. All contour plots in **Figure 7** are displayed after smoothing using a 2D Gaussian smoothing kernel with a standard deviation of 1. The curve fit of GEV-DV d_G as a function of study protocol, as shown in **Figure 8a-b**, was carried out using the Curve Fitting Toolbox in MATLAB R2018b, robustly by minimizing the least absolute residuals (LAR), using the Levenberg-Marquardt algorithm, and coefficient and function termination tolerances ('TolX' and 'TolFun') of 10^{-12}; all fits converged.

Data Availability

The data employed here are publicly available from the Human Connectome Project at https://db.humanconnectome.org, and include all *S500 Release subjects* resting state fMRI data[13,24,34,35].

Code Availability

The software package accompanying this article allows a user to perform the LPF-FD and GEV-DV volume censoring methods described herein, as well determine optimal volume censoring parameters for a given dataset size as per the formulae in **Table 2**. This code is publicly available under the terms of the GNU General Public License Version 3 on the

MathWorks File Exchange at https://www.mathworks.com/matlabcentral/fileexchange/73479-multiband_fmri_volume_censoring.

Acknowledgements

This work was funded by K01 MH 107763. J. C. W. was also supported by F30 MH 122136.

Mark Slifstein, Anissa Abi-Dargham, Joseph Schwartz, and Yuefan Deng provided helpful commentary on various aspects of the methods presented here. Zu Jie Zheng, Philip Tubiolo, Alexander Eichert, Eilon Silver-Frankel, and Bernie Hung-Wei Chen assisted with submitting jobs to the SeaWulf HCP cluster, and with generating figures. The Stony Brook University Institute for Advanced Computational Science (IACS) provided critical high-performance computing resources and technical assistance for the completion of this work.

Author Contributions

J.X.V.S. initially conceived of the project and J.C.W. additionally contributed to the design. Both authors contributed to the analysis and interpretation of data, as well as to the creation of new software. Both authors drafted and revised the manuscript.

Competing Interests Statement

The authors declare no competing financial interests.

References

1. https://www.nimh.nih.gov/about/strategic-planning-reports/strategic-research-priorities/index.shtml. Strategic Research Priorities Overview. (2017).

2. Ciric, R., *et al.* Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* **154**, 174-187 (2017).

3. Parkes, L., Fulcher, B., Yucel, M. & Fornito, A. An evaluation of the efficacy, reliability, and sensitivty of motion correction strategies for resting-state functional MRI. *NeuroImage* **171**, 415-436 (2018).

4. Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L. & Petersen, S.E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142-2154 (2012).

5. Power, J.D., *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320-341 (2014).

6. Power, J.D., Schlaggar, B.L. & Petersen, S.E. Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage* **105**, 536-551 (2015).

7. Satterthwaite, T.D., *et al.* An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage* **64**, 240-256 (2013).

8. Satterthwaite, T.D., *et al.* Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage* **60**, 623-632 (2012).

9. Van Dijk, K.R., Sabuncu, M.R. & Buckner, R.L. The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* **59**, 431-438 (2012).

10. Yan, C.G., *et al.* A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *Neuroimage* **76**, 183-201 (2013).

11. Power, J.D., *et al.* Customized head molds reduce motion during resting state fMRI scans. *Neuroimage* **189**, 141-149 (2019).

12. Moeller, S., *et al.* Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magn Reson Med* **63**, 1144-1153 (2010).

13. Ugurbil, K., *et al.* Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *NeuroImage* **80**, 80-104 (2013).

14. Miller, K.L., *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* **19**, 1523-1536 (2016).

15. Jernigan, T.L., Brown, S.A. & Coordinators, A.C. Introduction. *Dev Cogn Neurosci* **32**, 1-3 (2018).

16. Casey, B.J., *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cogn Neurosci* **32**, 43-54 (2018).

17. Volkow, N.D., *et al.* The conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev Cogn Neurosci* **32**, 4-7 (2018).

18. Hagler, D.J., Jr., *et al.* Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *Neuroimage*, 116091 (2019).

19. Burgess, G.C., *et al.* Evaluation of Denoising Strategies to Address Motion-Correlated Artifacts in Resting-State Functional Magnetic Resonance Imaging Data from the Human Connectome Project. *Brain Connect* **6**, 669-680 (2016).

20. Power, J.D., *et al.* Distinctions among real and apparent respiratory motions in human fMRI data. *Neuroimage* **201**, 116041 (2019).

21. Fair, D.A., *et al.* Correction of respiratory artifacts in MRI head motion estimates. *bioRxiv* (2018).

22. Power, J.D., *et al.* Characteristics of respiratory measures in young adults scanned at rest, including systematic changes and "missed" deep breaths. *Neuroimage* **204**, 116234 (2019).

23. Smyser, C.D., *et al.* Longitudinal analysis of neural network development in preterm infants. *Cereb Cortex* **20**, 2852-2862 (2010).

24. Van Essen, D.C., *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62-79 (2013).

25. Power, J.D., *et al.* Functional network organization of the human brain. *Neuron* **72**, 665-678 (2011).

26. Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B. & Bandettini, P.A. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *NeuroImage* **44**, 893-905 (2009).

27. Siegel, J.S., *et al.* Data Quality Influences Observed Links Between Functional Connectivity and Behavior. *Cereb Cortex* **27**, 4492-4502 (2017).

28. Power, J.D., Plitt, M., Kundu, P., Bandettini, P.A. & Martin, A. Temporal interpolation alters motion in fMRI scans: Magnitudes and consequences for artifact detection. *PLoS One* **12**, e0182939 (2017).

29. Power, J.D., *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* **115**, E2105-E2114 (2018).

30. Muschelli, J., *et al.* Reduction of motion-related artifacts in resting state fMRI using aCompCor. *Neuroimage* **96**, 22-35 (2014).

31. Ciric, R., *et al.* Mitigating head motion artifact in functional connectivity MRI. *Nat Protoc* **13**, 2801-2826 (2018).

32. Bassett, D.S., Xia, C.H. & Satterthwaite, T.D. Understanding the Emergence of Neuropsychiatric Disorders With Network Neuroscience. *Biol Psychiatry Cogn Neurosci Neuroimaging* **3**, 742-753 (2018).

33. Caballero-Gaudes, C. & Reynolds, R.C. Methods for cleaning the BOLD fMRI signal. *Neuroimage* **154**, 128-149 (2017).

34. Glasser, M.F., *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105-124 (2013).

35. Smith, S.M., *et al.* Resting-state fMRI in the Human Connectome Project. *Neuroimage* **80**, 144-168 (2013).

36. Prescott, P. & Walden, A.T. Maximum likelihood estimation of the parameters of the generalized extreme-value distrubtion. *Biometrika* **67**, 723-724 (1980).

37. Efron, B. & Tibshirani, R. *An introduction to the bootstrap* (Chapman & Hall, New York, 1993).

38. Wackerly, D.D., Mendenhall, W. & Scheaffer, R.L. *Mathematical statistics with applications* (Thomson Brooks/Cole, Belmont, CA, 2008).

39. Shi, Y. & Eberhart, R. A modified particle swarm optimizer. in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)* 69-73 (1998).

40. Kennedy, J. & Eberhart, R. Particle swarm optimization. in *Proceedings of ICNN'95 - International Conference on Neural Networks* 1942-1948 vol.1944 (1995).

41. Mezura-Montes, E. & Coello Coello, C.A. Constraint-handling in nature-inspired numerical optimization: Past, present and future. *Swarm and Evolutionary Computation* **1**, 173-194 (2011).

42. Ingber, L. Adaptive simulated annealing (ASA): Lessons learned. *Control Cybernetics* **25**, 33-54 (1996).

43. Kirkpatrick, S., Gelatt, C.D., Jr. & Vecchi, M.P. Optimization by simulated annealing. *Science* **220**, 671-680 (1983).

44. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087-1092 (1953).

45. Davidon, W. Variable Metric Method for Minimization. *SIAM Journal on Optimization* **1**, 1-17 (1991).

46. Conn, A.R., Gould, N. & Toint, P.L. A globally convergent Lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds. *Math. Comput.* **66**, 261-288 (1997).

47. Audet, C. & Dennis, J. Analysis of Generalized Pattern Searches. *SIAM Journal on Optimization* **13**, 889-903 (2002).

48. Lewis, R., Shepherd, A. & Torczon, V. Implementing Generating Set Search Methods for Linearly Constrained Minimization. *SIAM Journal on Scientific Computing* **29**, 2507-2530 (2007).

49. Abramson, M., Audet, C., Dennis, J. & Digabel, S. OrthoMADS: A Deterministic MADS Instance with Orthogonal Directions. *SIAM Journal on Optimization* **20**, 948-966 (2009).

50. Kolda, T., Lewis, R. & Torczon, V. Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods. *SIAM Review* **45**, 385-482 (2003).