Advancing motion denoising of multiband resting state functional connectivity fMRI data

John C. Williams[1,2], Philip N. Tubiolo[2], Jacob R. Luceno[1], and Jared X. Van Snellenberg[1,2,3,4,5,*]

Affiliations

[1]Department of Psychiatry and Behavioral Health, Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, 11794;

[2]Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY, 11794;

[3]Department of Psychology, Stony Brook University, Stony Brook, NY, 11794;

[4]Division of Translational Imaging, New York State Psychiatric Institute, New York, NY, 10032;

[5]Department of Psychiatry, Columbia University Medical Center, New York, NY, 10032.

*To whom correspondence may be addressed. **Email:** Jared.VanSnellenberg@stonybrookmedicine.edu.

Address: 101 Nicolls Rd., Health Sciences Center T10-087J, Stony Brook, NY 11794.

Abstract

Multiband-accelerated fMRI provides dramatically improved temporal and spatial resolution of resting state functional connectivity (RSFC) studies of the human brain, but poses unique challenges for denoising of subject motion induced data artifacts, a major confound in RSFC research. We comprehensively evaluated existing and novel approaches to volume censoring-based motion denoising in the Human Connectome Project dataset. We show that assumptions underlying common metrics for evaluating motion denoising pipelines, especially those based on quality control-functional connectivity (QC-FC) correlations and differences between high- and low-motion participants, are problematic, making these criteria inappropriate for quantifying pipeline performance. We further develop two new quantitative metrics that are free from these issues and demonstrate their use as benchmarks for comparing volume censoring methods. Finally, we develop rigorous, quantitative methods for determining optimal censoring thresholds and provide straightforward recommendations and code for all investigators to apply this optimized approach to their own RSFC datasets.

Introduction

The study of resting state functional connectivity (RSFC) with functional Magnetic Resonance Imaging (fMRI) 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, it is widely recognized that participant motion represents a serious confound in RSFC research, and that aggressive steps must be taken to minimize its impact (*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 (multiband) acceleration of fMRI sequences (*12, 13*) has gained prominence, 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 fMRI data. However, data acquired with sub-second repetition times (TRs), as enabled by multiband acceleration, appear to contain not just traditionally recognized markers of motion artifact (e.g., net elevations in observed RSFC correlations, with both spatially-dependent and spatially-independent components), but also novel artifacts not observed in traditional, higher-TR single-band data. Fair and colleagues (*19*) and Power and colleagues (*20*) evaluated data from the ABCD and HCP studies (respectively), and concluded that both true respiratory motion (most pronounced as changes in pitch, as well as in anterior-posterior and vertical position) and "false" pseudomotion (factitious head motion, observed primarily in the phase-encode direction as a result of tissue changes due to lung expansion impacting the B_0 field) are evident in multiband fMRI due to the higher sampling rate of these sequences relative to traditional single-band fMRI, findings which parallel our own observations (*21-23*; see **Figure S1** for an example).

The discovery of novel forms of motion and pseudomotion artifact (hereafter, "motion artifact," except where this distinction is of note) in fast-TR multiband data underscores a critical need to comprehensively evaluate methods for characterizing and removing motion artifact in these data. Previously, Burgess and colleagues (*24*) evaluated volume censoring (a widely used motion artifact correction approach) at a single threshold, independent components analysis-based denoising (ICA-FIX;

25, 26), and global signal regression (GSR, via mean greyordinate timeseries regression; MGTR) in the HCP multiband dataset and concluded that a combination of ICA-FIX and MGTR showed the best denoising performance, although detectable motion artifact remained in the data, and any prescription for widespread use of GSR/MGTR remains controversial (*27-32*). Moreover, while volume censoring reduced spatially specific artifacts, its performance was significantly impaired in the context of multiband data (*24*), at least in part due to the aforementioned presence of respiratory motion (*20*). However, notch filtering of motion estimates prior to calculation of the subject-level quality control (QC) measure framewise displacement (FD) has been shown in other work to lead to improved performance of volume censoring pipelines (*19, 20*), although this work did not attempt to modify measures of the temporal derivative root-mean-squared variance over voxels (DV), another commonly used measure.

Thus, while it seems clear that volume censoring behaves differently in fast-TR multiband datasets than it does in singleband datasets, and some improvement in volume censoring methods has been made over existing singleband methods (*19, 20*), we suspected that substantive improvement in these methods may still be possible. For example, while notch filtering of MPs improves FD-based volume censoring (*19, 20*), such filters have not been attempted for DV-based censoring (widely used alongside FD censoring), and filtering changes the magnitude of FD values, raising the possibility that common FD censoring thresholds (e.g., of .2 or .5 mm; see *5*) may no longer be appropriate.

Consequently, in this paper we evaluate several volume censoring approaches (FD and DV-based censoring, with or without filtering and GSR) across the *full range* of potential thresholds using common metrics for assessing the motion artifact remaining in a dataset, to rigorously, and quantitatively, determine optimal methods and threshold for censoring multiband RSFC data. Consistent with conclusions drawn by others (*19, 20*), we hypothesized that denoising approaches that ignore motion due to respiration (and potentially other high-frequency noise sources that would be removed by standard .01 - .1 Hz bandpass filtering of RSFC data) and censor only those volumes that exhibit irregular and sudden (and thus predominantly low frequency) motion might perform as well or better than existing methods. Thus, we developed a volume censoring approach that low-pass filters at 0.2 Hz prior to calculation of FD and DV (which we term LPF-FD and LPF-DV, respectively), a frequency low enough to remove nearly all respiration-related motion while remaining at least double the upper frequency band of standard RSFC

band-pass filters. 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 (as has been previously recognized; see *33, 34*), such that an apparently reasonable cutoff for one run could potentially remove all data from another (see **Figure S2A-B**). Consequently, we also tested an adaptive thresholding method (demonstrated in **Figure S2C**) that fits a generalized extreme value (GEV) distribution to the LPF-DV values within each run separately and rejects the upper tail (i.e., outliers) of that distribution on a per-run basis (GEV-DV censoring).

Thus, we evaluate each censoring approach (LPF-FD, LPF-DV, and GEV-DV) against standard volume censoring methods (*5*) that have performed well in recent evaluations of motion denoising strategies (*2, 3*), using the HCP S500 release (*35*). We initially set out to employ a set of dataset-level QC metrics (hereafter, DQMs) that have been employed in prior work (*2-10, 24, 36, 37*), but we observed substantive issues with DQMs that depend on associations between subject-level QC metrics (such as mean or median FD; hereafter, SQMs) and the observed magnitude of RSFC correlations (e.g., so-called QC-FC correlations). These issues, which we detail in *Results and Discussion*, led us to develop a novel DQM that quantifies a common graphical approach to quality assessment (*4, 5*) while remaining agnostic to the particular spatial characteristics of artifact in the data.

Next, we developed and evaluated a quantitative, empirical, method for determining the optimal value of censoring thresholds for removing motion artifacts from RSFC data by optimizing the trade-off between reducing motion-induced bias and increasing variance (and loss of power) that results from data removal due to censoring, which we call ΔMSE-RSFC. This approach uses a bias-variance decomposition of mean squared error (MSE) of RSFC correlations to estimate the *total* error in a sample, including 1) motion-induced bias in sample mean RSFC correlations; 2) variance resulting from both true variability between subjects and the unequal distribution of motion, and thus motion artifact, across individuals and runs; and 3) sampling error within each run. Next, we applied a global optimization algorithm to this measure to determine optimal parameters for simultaneous FD- and DV-based censoring in the HCP500 dataset by minimizing ΔMSE-RSFC.

Finally, we developed a method to generalize these results to multiband datasets of nearly any size, so they can be employed outside the context of the HCP dataset investigated here, and provide a

straightforward formula and code for use by investigators to estimate optimal thresholds for their datasets. These results are appropriate for use by investigators to determine censoring thresholds prior to data collection or analysis, which is important for reducing bias (and "experimenter degrees of freedom") and necessary to allow the inclusion of precise denoising methodology in study preregistration (*38*), improving the reliability and reproducibility of RSFC fMRI studies.

Results and Discussion

Standard Quality Control Benchmarks Across a Full Range of Censoring Thresholds

First, we calculated widely used QC metrics for evaluating the success of motion denoising of RSFC datasets (i.e., DQMs) across a comprehensive range of potential censoring thresholds, ranging from no censoring to thresholds that approach removal of all data in the dataset. Although extremely computationally intensive, we view this as a critical step in establishing which DQMs are best suited for determining the success or failure of volume censoring techniques, or for optimizing such techniques.

We show several commonly-used DQMs in **Figure 1**, calculated after either standard FD or LPF-FD based volume censoring, and presented as a function of the percentage of volumes in the entire dataset that are removed as a result of censoring (to allow direct comparison between methods despite differing threshold values), with identified "optima" in **Table S1**. Results for DV-based measures are shown in **Figure S3** (GEV-DV) and **Figure S4** (Standard DV and LPF-DV), with "optima" shown in **Table S2** and **Table S3**. Taken together, the results shown in **Figure 1** raise serious questions as to whether several of these DQMs are appropriate benchmarks for evaluating the denoising of RSFC datasets. For example, each of: a) the median absolute value of all QC-FC correlations (**Figure 1A;** see *2*), b) the proportion of ROI pairs with statistically significant QC-FC correlations (**Figure 1B;** see *2, 39*), and c) the proportion of ROI pairs showing group differences between high- and low-motion terciles of the dataset (high-low null rejection rate, terciles determined by mFD; **Figure 1D;** see *5, 7, 9, 24, 37*) show nearly immediate increases, rather than the expected decreases, as censoring becomes more aggressive for both standard FD and LPF-FD when GSR is employed. A similar effect is seen with standard FD without

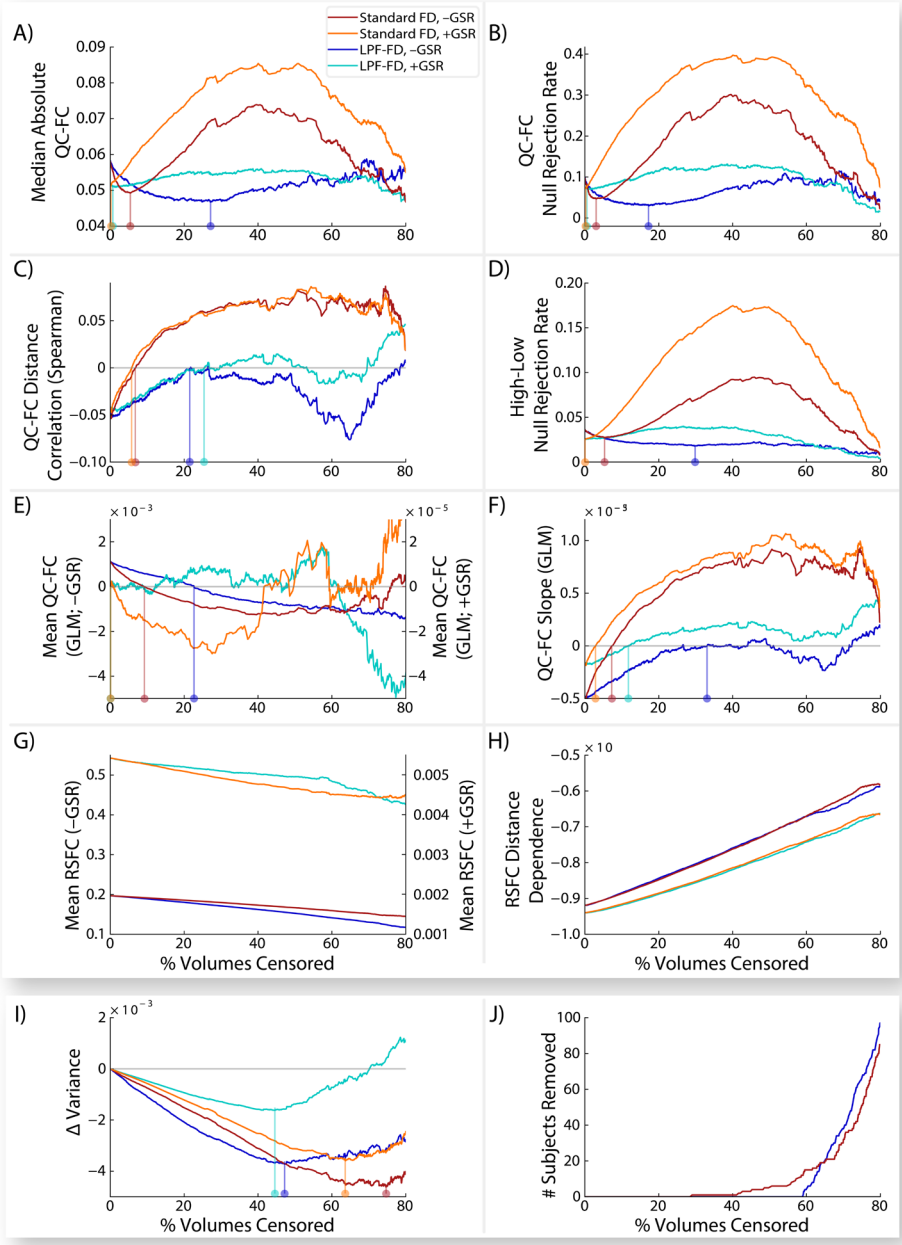

Figure 1. Dataset quality control metrics (DQMs) utilized in past literature as a function of the percent of volumes censored in a subset of the HCP500 dataset (*Dataset 1a*; n = 475). DQMs calculated after volume censoring using either standard framewise displacement (FD) or low-pass-filtered FD (LPF-FD). **A)** Median absolute QC-FC across ROI pairs, **B)** proportion of ROI pairs with significant QC-FC correlations (null rejection rate), **C)** correlation of QC-FC with the distance between ROI pairs, and **D)** proportion of ROI pairs showing significant differences between high-motion and low-motion participants. Also shown are DQMs calculated using a two-step generalized linear model (GLM), including: **E)** the mean of QC-FC correlations across the dataset, **F)** the distance-dependence (slope) of QC-FC correlations, **G)** the mean of RSFC correlations across the dataset, and **H)** the mean distance-dependence (slope) of RSFC correlations, as well as **I)** the change in between-subjects variance due to censoring, and **J)** the number of subjects removed due to volume censoring. Vertical lines show potential "optima" according to each DQM when identifiable. Results from analyses without and with global signal regression (GSR) are respectively denoted –GSR and +GSR. Data in panel **J** are identical for analyses with and without GSR.

GSR, but after a clearer “initial dip”. While this dip may suggest an optima, the properties of the data that drive the dramatic rise in these metrics after this optima are unclear. At face value, these results suggest that with GSR virtually no censoring need be performed, and that without GSR, LPF-FD achieves better data quality than standard FD, but at the cost of censoring 4-5 times as much data. However, as we argue in more detail below, our view is that these findings instead reflect fundamental issues with this set of DQMs, and that alternative DQMs should be employed instead.

The rank-order correlation between QC-FC correlations and ROI pair distance (**Figure 1C;** see *2, 3, 4, 7, 36*) raises a distinct set of concerns. While standard FD censoring (with or without GSR) results in an apparent rapid “improvement” to an absence of distance-dependence (i.e., the complete removal of the typically observed effect that short-distance RSFC correlations have a stronger relationship to QC measures such as mFD than do long-distance RSFC correlations), further censoring with standard FD reverses the distance dependence effect. This challenges the notion that standard FD censoring is simply removing the distance-dependent effect of motion artifact, because in that case a reverse-distance-dependence effect should not be possible—instead, distance-dependence should reach or approach 0 and then plateau. It is unclear how this reversal could occur, or how it should be interpreted. While LPF-FD requires substantially more censoring to reach a QC-FC distance correlation of 0, the fact that it does *not* result in QC-FC distance correlations greater than 0 may indicate that it is a preferable censoring method, although censoring past approximately 50% of frames removed causes this metric to worsen again. Regardless, the behavior of this metric across these censoring approaches undermines clear interpretation of results employing this DQM.

We also evaluated four DQMs from a generalized linear model (GLM) approach developed by Burgess and colleagues (*24*) that estimates a) mean RSFC **(Figure 1G),** b) the mean distance-dependence of RSFC correlations (**Figure 1H;** see *4, 5, 6, 8*)**,** c) the mean QC-FC value over all ROI pair correlations **(Figure 1E;** see *2, 5-7, 36)*, and d) the slope of QC-FC and distance **(Figure 1F;** see *2, 4, 8*). Results from the slope of QC-FC and distance almost precisely mirror the correlation between QC-FC and distance, as discussed above. With GSR, mean QC-FC begins near zero (the theoretically optimal value) and only exhibits small fluctuations with additional censoring. Without GSR, mean QC-FC shows steady reductions with more censoring, but these values become negative, raising a set of questions

parallel to those outlined above regarding the QC-FC distance correlation shown in **Figure 1C**; namely, it is unclear how to interpret distance-dependence measures when increasingly aggressive censoring pushes the metric past its theoretically optimal value into a “reversed” or “flipped” state from the raw, uncensored, data.

However, the two measures based directly on RSFC correlations, rather than on their association with mFD (as in the QC-FC DQMs), suggest that regardless of censoring method increasingly stringent censoring results in continuous reductions in both distance-dependence and the magnitude of RSFC correlations. Unlike the other DQMs, this provides a clearly interpretable result, although it leads to the unfortunate conclusion that even the smallest motions have a measurable impact on RSFC, and that this influence can only be removed by censoring until the entire dataset has been discarded—clearly, not a practicable solution to the issue of motion in RSFC.

Finally, **Figure 1I-J** shows the average (across ROIs) change in between-subject variance in the dataset as a result of FD-based censoring (and for DV-based censoring in **Figure S5** and **Figure S6**), as well as the number of participants removed from the dataset at each threshold. This demonstrates that the behavior of the DQMs shown in **Figure 1A-F** cannot be attributed to removal of subjects causing changes to the dataset. In addition, increasingly aggressive censoring leads to marked reductions in between-subjects variance, demonstrating that the removal of the highest FD volumes from the dataset makes RSFC values across subjects more similar. Moreover, this effect continues well past the putatively ‘optimal’ censoring thresholds identified by every DQM evaluated here. Thus, there are clear data-quality impacts on RSFC correlations, at least in terms of reducing motion-associated subject-level variability in RSFC, as a result of volume censoring that occur at much higher censoring thresholds than are suggested by commonly used DQMs, or than is widely appreciated in the existing literature.

Exploration of Confounds in mFD-Based Dataset-QC Metrics

One possibility raised by the above results is that there is a fundamental problem with DQMs based on QC-FC relationships. Every DQM reviewed above that depends on mFD, or another SQM (e.g., proportion of volumes censored, see **Figure S7**, **Figure S8**, and **Figure S9**), behaves in a way that is not clearly interpretable, while every DQM that does not depend on subject-level QC metrics shows

monotonic changes in the expected direction with additional censoring, and never exceeds a theoretically optimal value. Consequently, we conducted follow-up analyses to examine whether one or more confounds may impact DQMs that rely on SQMs.

We hypothesized that these DQMs could be confounded by true differences in RSFC correlations between high- and low-motion participants. That is, these DQMs rest on the assumption that the true relationship between mFD and RSFC in a motion-free dataset is 0. However, there could be true differences in RSFC between high- and low-motion individuals that are unrelated to motion artifact; that is, a "third variable" confound. This is consistent with reports that RSFC differences have been observed between high- and low-motion participants, even when only considering low-motion scans (*40*), and that the quantity and quality of participant motion is significantly associated with participant demographic characteristics (*41, 42*). Such confounds could call into question not just the assumption that the observed value of QC-FC correlations (or number of significant differences between high- and low-motion subjects) in a dataset should optimally be zero, but also that *any* target value for such metrics can be known a priori.

In light of these concerns, we decided to test whether two likely candidates for a third variable that could impact both RSFC correlations and participant motion may be impacting the QC-FC null rejection rate (the DQM reported in **Figure 1B**). **Figure 2** shows the observed proportion of RSFC correlations that significantly differ between high- and low-motion participants (upper and lower tercile on median FD, respectively) at 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 2** 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 ***Materials and 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

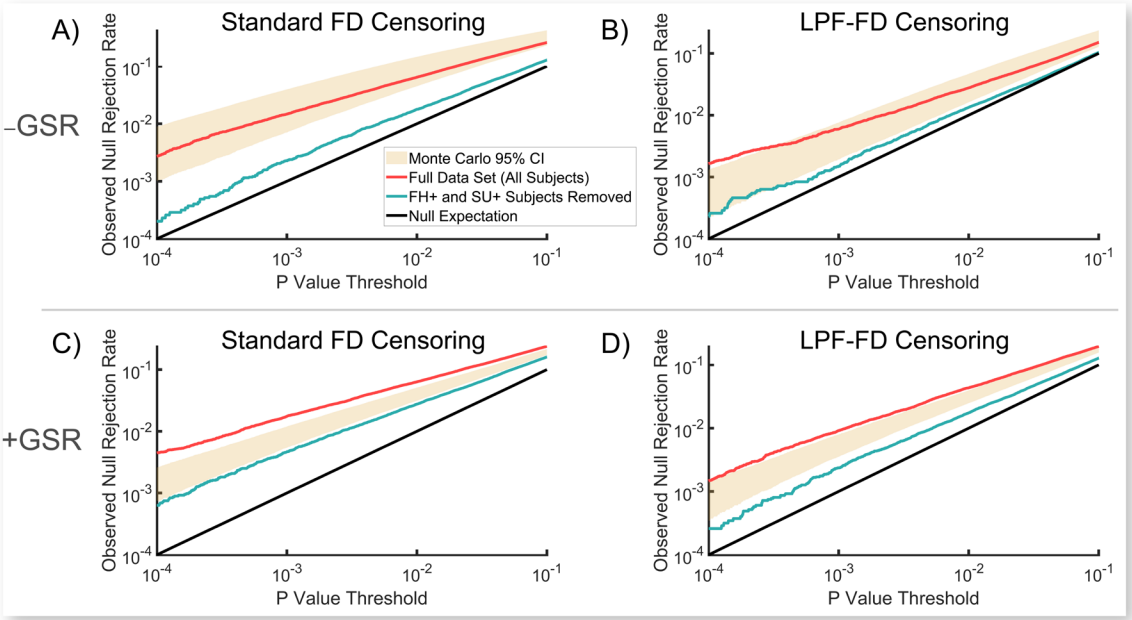

Figure 2. Proportion of significant RSFC correlations between high- and low-motion participants at various uncorrected P value thresholds in the HCP500 dataset (n = 501; *Dataset 1)*. The proportion observed for the full dataset (red) and after removing subjects who tested positive for drug use or elevated blood alcohol content on the day of a scan session (SU+) or have a family history (FH+) of a psychiatric or neurological disorder (blue-green; n = 315; *Dataset 1b*) are shown. 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 for analyses with standard volume censoring (left) and LPF-FD volume censoring (right), without global signal regression (GSR; –GSR, top) and with GSR (+GSR, bottom). Censoring thresholds selected to ensure an equal number of volumes removed across methods: **A)** FD > 0.1475 mm **B)** LPF-FD > 0.0318 mm **C)** FD > 0.155 mm **D)** LPF-FD > 0.0337 mm.

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 (*41*).

One potential critique of this analysis is that matching FH+ and SU+ subjects to FH- and SU- subjects on overall motion (i.e., mFD) may fail to capture important differences in the *type* of motion occurring in these two groups; that is, even if mFD is equivalent in two individuals, one individual may, e.g., exhibit relatively steady motion throughout the scan, while another exhibits lower motion through

most of the scan but has several large individual motions that result in equivalent mFD values between these individuals. In order to evaluate this possibility, we developed a matching algorithm that finds pairs of subjects who have maximal overlap in the empirical cumulative density functions (ECDFs) of the derivatives of their motion parameter (MP) traces, based on calculation of the Cramér–von Mises criterion (*43*). However, because this approach uses one-to-one matching of subjects, it is not amenable to the analysis in **Figure 2**. Instead, we employed a contingency table analysis to determine whether the high-low null rejection rate differed after removing FH+ versus FH- subjects. **Figure 3** demonstrates that the high-low null rejection rate is significantly higher in samples with FH+ remaining than in those with all FH+ removed across all P-value thresholds examined (null rejection rates shown in **Figure S10**). We repeated these analyses with FH- subjects instead matched by 1) averaging the Cramér–von Mises criterion calculated on the ECDFs of the derivatives of MPs together with the Cramér–von Mises criterion calculated on the ECDFs of DV traces (**Figure S11** and **Figure S12**), and 2) the means of the absolute

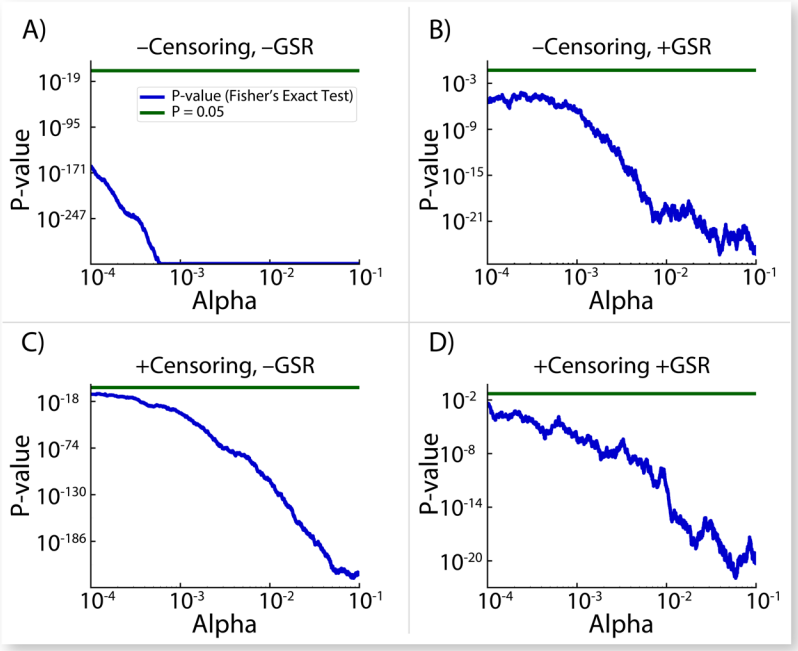

Figure 3. Analysis of the change in number of significant differences in RSFC between high- and low-motion subjects as a function of whether subjects with a family history of a psychiatric or neurological disorder (FH+) or FH- subjects were removed from the dataset. FH- subjects were matched one-to-one with 161 FH+ subjects on the empirical cumulative density functions (ECDFs) of the derivatives of their motion parameter traces in a subsample of unrelated subjects from the HCP1200 Subjects Release (total n = 403; *Dataset 2*). In both groups with either FH+ or FH- removed, n = 242. Significance established using Fisher's exact test (blue lines); green lines indicate p = 0.05 (the nominal alpha level). Analyses without GSR are denoted by –GSR (left), and those with GSR are denoted by +GSR (right); analyses with volume censoring are denoted by –Censoring (top), and those with volume censoring are denoted by +Censoring (bottom).

values of the derivates of MP traces (**Figure S13** and **Figure S14**), as described in *Subject Matching Algorithms* (see ***Materials and Methods***).

Finally, we used a matched FH+ and FH- dataset from the above analyses and measured the effect of accounting for FH+ group membership in a partial correlation when calculating 1) median absolute QC-FC, 2) proportion of significant QC-FC correlations, and 3) QC-FC slope (Spearman's), and in a parallel analysis where FH+ was a nuisance regressor in the GLM-based method (*24*) used to determine mean QC-FC and the association between QC-FC and ROI pair distance. Results for analyses using uncensored data with no nuisance regressors (to show these results do not depend on effects of motion denoising pipelines) are reported in **Table 1.** We repeated these analyses using 161 FH- participants matched instead by the ECDFs of both MP derivative and DV traces **(Table S4)**, as well as by the mean absolute value of MP derivatives **(Table S5)**. Regardless of the method used to match subjects based on motion, significant differences are observed when accounting for group membership in the majority of DQMs, suggesting that these metrics are critically confounded.

Table 1. Permutation test results showing the impact of modeling family history status on dataset quality metrics (DQMs) with subjects matched on motion parameter (MP) derivative distributions.

Dataset Quality Metric (DQM)	Value	Null Expectation [CIs]	Difference (% Difference)	P-Value
QC-FC Slope (Spearman's ρ)	**-1.27×10^{-1}**	**-1.39×10^{-1} [-1.44×10^{-1}, -1.34×10^{-1}]**	**1.16×10^{-2} (-8.35%)**	**0.00240****
QC-FC Null Rejection Rate	**7.17×10^{-3}**	**9.91×10^{-3} [8.41×10^{-3}, 1.16×10^{-2}]**	**-2.74×10^{-3} (-27.62%)**	**0.00560****
Median Absolute QC-FC	4.75×10^{-2}	4.78×10^{-2} [4.74×10^{-2}, 4.84×10^{-2}]	-3.64×10^{-4} (-0.76%)	0.0708
QC-FC Slope (GLM; QC = mFD)	**-8.09×10^{-6}**	**-9.55×10^{-6} [-1.47×10^{-5}, -8.64×10^{-6}]**	**1.46×10^{-6} (-15.32%)**	**0.0128****
Mean QC-FC (GLM; QC = mFD)	**-9.18×10^{-4}**	**-3.94×10^{-4} [-6.98×10^{-4}, 2.07×10^{-5}]**	**-5.24×10^{-4} (133.12%)**	**0.00780****

Bolding and double asterisks (**) denote significance after false discovery rate (FDR) correction with α = 0.05. QC-FC = Quality Control - Functional Connectivity; mFD = Mean Framewise Displacement. "GLM" (Generalized Linear Model) denotes statistics generated using the GLM-based data quality metrics (DQMs) developed by Burgess and associates, with mean FD used as the subject QC metric. Null expectation and confidence intervals (CIs) determined as the median of a null distribution generated using 10,000 permutations of family history positive (FH+) group membership assignments. Results shown from a subset of the HCP1200 dataset (n = 242; Dataset 2a) with 161 FH+ and 161 FH- subjects, matched by the empirical cumulative density functions (ECDFs) of the derivatives of their motion parameter (MP) traces.

Thus, DQMs that depend on relationships between RSFC and an SQM (i.e., QC-FC measures) are clearly impacted by unmeasured third variables. Moreover, although we demonstrate here that FH status acts on both measures of participant motion and on RSFC values, thus confounding any DQM that depends on the relationship between motion and RSFC, we cannot rule out that other, unmeasured, individual differences could also have this impact. Thus, unless *all* participant characteristics that impact both motion and true RSFC correlations can be identified and modeled, DQMs that implicitly assume that such confounds do not exist should not be employed to evaluate the effectiveness of motion denoising strategies on RSFC data.

Alternative Methods for Comparing Standard and LPF-Based Censoring

Because of the issues described above with established DQMs, we sought alternative methods of evaluating the impact of denoising on RSFC, beginning with a visual comparison approach employed in prior work (*4, 5*). Changes in RSFC correlations resulting from motion-targeted volume censoring are shown for all ROI pairs (*44*) in **Figure 4** (red dots). These changes are compared to RSFC changes resulting from removal of an identical number of randomly selected volumes within each run (black dots).

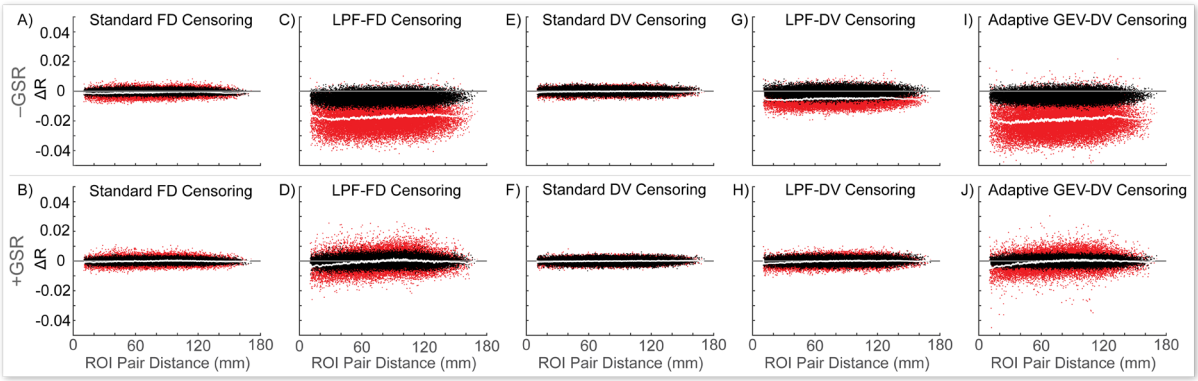

Figure 4. Changes in pairwise ROI RSFC correlations resulting from volume censoring. Results shown after censoring using standard FD **(A,B)** and standard DV **(E,F)**, LPF-FD **(C,D)** and LPF-DV **(G,H)**, and adaptive GEV-DV **(I,J)**, plotted against the Euclidean distance between ROIs on the X-axis, in the HCP500 dataset (n = 501; *Dataset 1*). 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; –GSR, top) and with GSR (+GSR, bottom). Censoring thresholds were selected such that an equal amount of data was removed by each method (56.15% of volumes in analyses without GSR, 40.06% of volumes in analyses with GSR) as follows. Without GSR: **A)** FD > 0.1475, **C)** LPF-FD > 0.0318, **E)** DV > 61, **G)** LPF-DV > 10.75, and **I)** GEV-DV d_G = 1.16. With GSR: **B)** FD > 0.1550, **D)** LPF-FD > 0.0337, **G)** DV > 63, **H)** LPF-DV > 11.25, and **J)** GEV-DV d_G = 1.36.

Thresholds were selected such that an equivalent number of volumes were removed by each evaluated method, to allow an apples-to-apples comparison between censoring methods.

Standard (unfiltered) censoring methods resulted in minimal changes in RSFC compared to random censoring of an equivalent number of frames within each run, consistent with the view that this results in the targeting of respiratory motion and pseudomotion in high-TR datasets (*20*), signal effects that, because of their relatively high frequency, should be effectively removed from the data by standard bandpass filtering of RSFC data, and thus are not ideal targets for volume censoring. Standard methods were outperformed by LPF-based censoring (**Figure 4A-H**), and GEV-DV censoring produced an even greater change in observed RSFC than LPF-DV censoring (**Figure 4G-J**). Finally, in analyses without GSR (**Figure 4, top row**), 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, 31*), 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. 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 S15**).

Next, we attempted to quantify the visual comparison in **Figure 4** to evaluate it across a range of censoring parameters, as we did for other DQMs**.** Thus, we developed a new DQM based on this visual comparison: we compute the mean absolute value of the within-subject change in RSFC correlations relative to randomly removing an equivalent number of volumes within each run, across all ROI pairs, and across all subjects in the sample, thus effectively quantifying the spread of red dots away from the black dots in **Figure 4**. Observed between-method differences in this metric should be specifically associated with differences in targeting of BOLD signal fluctuations resulting from head motion (or pseudomotion), rather than any other source. We term this measure MAC-RSFC (Mean Absolute Change in Resting State Functional Connectivity). **Figure 5,** which summarizes over 6.95×10^{15} partial correlations, shows that 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 of possible censoring thresholds (from 0% to nearly 100% of data removed), the LPF-based methods we propose here significantly (given the non-overlap of 95% confidence intervals obtained from bootstrapping) outperformed standard censoring methods. In addition, GEV-DV outperformed LPF-DV, suggesting that, in line with our observations in **Figure S2**, **Figure 4G-J,** and **Figure S15I-L**, an adaptive thresholding method is preferable for handling the substantial differences in central tendency of DV measures across runs.

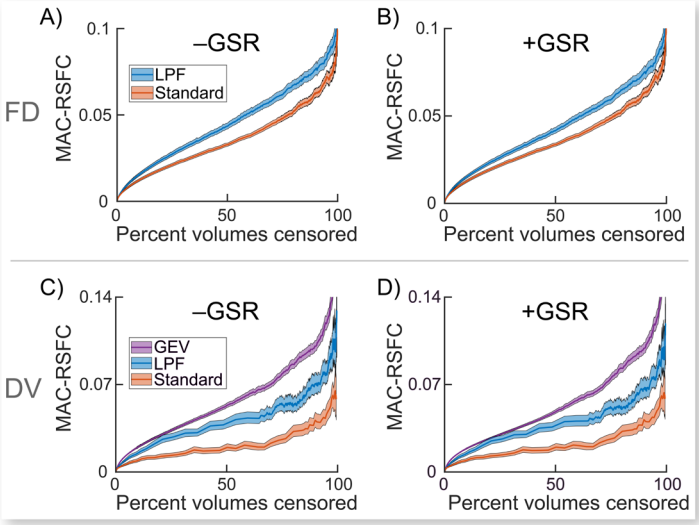

Figure 5. Mean Absolute Change in Resting State Functional Connectivity (MAC-RSFC) and 95% confidence intervals (CIs) as a function of the percent of volumes censored in the HCP500 dataset (n = 501; *Dataset 1*). Results shown after volume censoring using standard FD and LPF-FD **(A,B)** and standard 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 global signal regression (GSR; –GSR, left) and with GSR (+GSR; right).

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 (*19, 20*). 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, unfiltered, approaches to FD-based censoring on MAC-RSFC, they perform more poorly than LPF methods (**Figure S16)**. Thus, LPF-FD and GEV-DV volume censoring appears to be an across-the-board superior method of volume censoring of multiband data, no matter how aggressive a threshold is employed.

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

It has been noted that, irrespective of the censoring method used, increasingly strict thresholds result in continuous 'improvements' in data quality (*5, 6, 45*), an observation that is also borne out by all

DQMs evaluated here that do *not* depend on SQMs such as mFD (**Figure 5** and **Figure 1E-F**). This raises a critical challenge as to how to balance the tradeoff between the benefits of additional denoising and the costs of discarding additional data. Presently, such thresholds are typically selected from previously established thresholds (e.g. FD > 0.2 mm or FD > 0.5 mm in singleband data), or are sometimes adapted for particular datasets through qualitative visual inspections of a variety of DQMs. However, such approaches do not rest on a rigorous quantitative optimization. To our knowledge, the only formalized procedure for quantitatively selecting a censoring threshold in the literature involves calculating the correlation between RSFC estimates from a sliding window and the lowest-motion sliding windows in a given subject, and binning these estimates by the maximum FD value observed in the window (*5, 19*). The correlation of each window to the reference (low-motion) windows is then compared to randomly-ordered (permuted) data, and a threshold is selected such that sliding windows with higher FD values than the threshold show a significantly lower correlation to the reference data than is expected by chance.

Unfortunately, this approach has multiple drawbacks that make it unsuitable for selecting an optimal censoring threshold, either within a single study or in the literature more broadly. First, although threshold selection is based on a null-hypothesis statistical test (NHST), it is a misapplication of NHST to an optimization problem. There is no a priori reason why P = 0.05 should lead to an optimal censoring parameter, as compared to any other P value. Indeed, one could argue that the top 5th percentile, rather than the bottom, should be used (i.e., P = 0.95, rather than P = 0.05, in the original 1-tailed formulation of the test), which would reflect statistically significant evidence that a given time-window is *more similar* to the most motion-free data in a single subject than is an average (and thus more heavily motion-contaminated) time-window. Alternatively, the 50th percentile (P = 0.50) could be used as a cutoff, indicating simply that a time-window is more similar to the lowest-motion data than is average randomly-ordered data. Thus, while this method is quantitative, the choice of cutoff remains arbitrary. Moreover, this approach has the additional drawback of identifying different thresholds for different subjects, and will paradoxically select more lenient thresholds for higher motion subjects (because the null reference distribution generated from random timepoints should have more contamination from high-motion data, the null distribution in these subjects should have less similarity to the lowest-motion data and will thus

set the P < 0.05 cutoff at a higher FD). Indeed, work using this approach in the ABCD dataset (see *19, Figure 9B*) suggests an FD cutoff anywhere between approximately 0.05 and 0.3 mm depending on whether one desires a censoring threshold below the P < 0.05 threshold for *all* subjects or a threshold below this threshold in only the highest-motion subject (or somewhere in between).

Thus, we developed a novel method of determining optimal volume censoring thresholds by balancing the tradeoff between reducing the impact of motion and other artifacts on RSFC and minimizing the loss of tDoF (and thus power) that occurs when data is discarded. 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 motion-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 novel DQM that measures overall improvement in RSFC estimates resulting from motion denoising by employing a mean-squared error

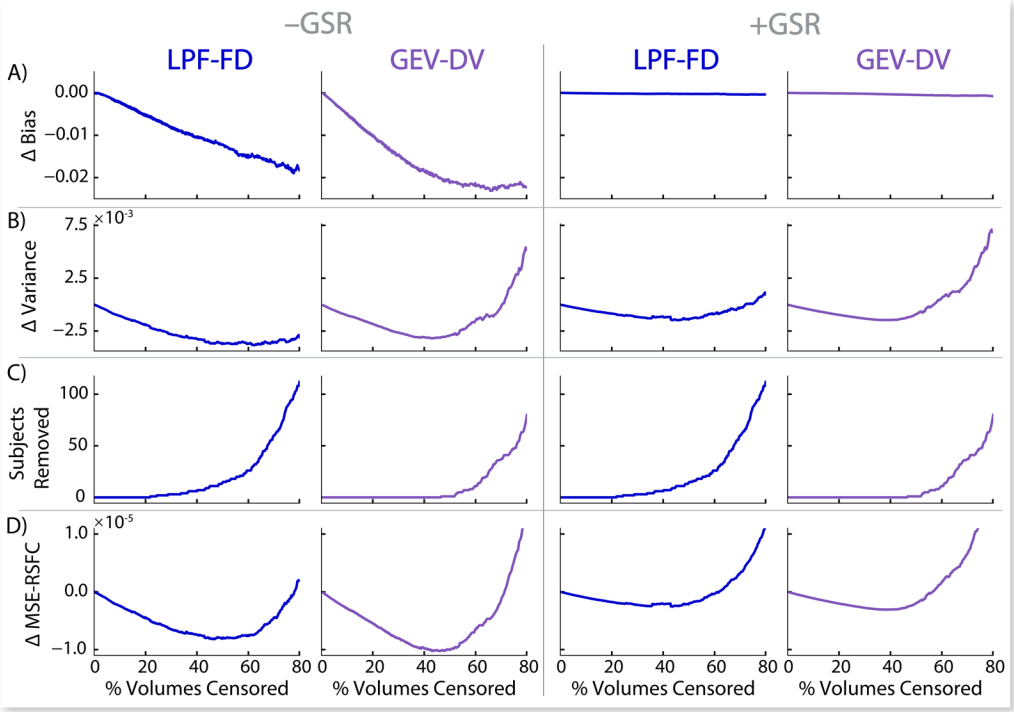

Figure 6. Change in sample statistics used to calculate ΔMSE-RSFC due to volume censoring in the HCP500 dataset (n = 501; *Dataset 1*). A) Mean change in sample average Z-transformed ROI pair correlations, over and above that due to random removal of an equivalent number of randomly selected volumes 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. Analyses without global signal regression (GSR) are denoted by –GSR (left); those with GSR are denoted by +GSR (right).

(MSE) calculation from a bias-variance decomposition (i.e., 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. **Figure 6** shows this approach for LPF-FD and GEV-DV censoring (used separately). **Figure 6A** demonstrates that increasingly aggressive censoring (nearly) continuously reduces motion-induced bias in RSFC, with significantly greater effects in analyses without GSR. Additionally, **Figure 6B** shows that both censoring methods produce a reduction in between-subjects variance that exceeds the increase caused by increasing sampling error until approximately 40-50% of volumes are removed. As variance is reduced before any subjects are removed (**Figure 6C**), the reduction in variance is not due to removal of high-motion participants; rather, it is due to the exclusion of volumes that were impacting individual RSFC correlation estimates in higher-motion subjects. Thus, contrary to our initial expectations, power to resolve sample-level RSFC correlations *increases* with more aggressive censoring, despite the loss of tDoF, as a result of reducing between-subject variance that exists because of motion effects on RSFC correlations. Finally, 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 against 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 method: 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 2**.

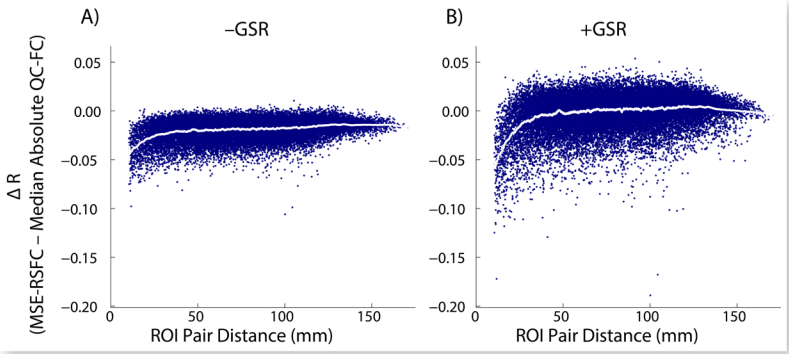

Figure 7. Change in ROI pair correlations resulting from additional censoring as indicated by ΔMSE-RSFC, over and above the optimal threshold indicated by Median Absolute QC-FC, in a subset of the HCP 500 dataset (n = 475; *Dataset 1a*). Results shown for analyses without global signal regression (GSR; –GSR, left) and with GSR (+GSR, right).

To allow for direct comparison between optimal volume censoring thresholds derived from ΔMSE-RSFC and those suggested by traditional DQMs (see above), we show in **Figure 7** the difference in RSFC correlations produced by LPF-FD censoring using ΔMSE-RSFC (demonstrated in **Figure S18**) over and above the optima suggested by minimizing median absolute QC-FC (*2*). This shows a general global reduction in RSFC correlations (without GSR), as well as a specific reduction in distance-dependent artifact. Equivalent plots for all other DQMs show similar results when using LPF-FD (**Figure S19**), GEV-DV (**Figure S20)**, standard FD (**Figure S21**), standard DV (**Figure S22**), and LPF-DV (**Figure S23**). That is, we show that censoring at the more aggressive thresholds provided by ΔMSE-RSFC optimization reduces the distance-dependence of RSFC correlations above and beyond the levels achieved by optimizing *any* known DQM that relies on SQMs such as mFD (i.e., QC-FC measures).

Table 2. Optimal parameters for volume censoring determined by minimizing *ΔMSE-RSFC.*

Optimal Volume Censoring Parameters for the HCP500 Dataset *(Dataset 1)*		
	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

Optimal Combined LPF-FD and GEV-DV Censoring Parameters as a Function of Dataset Size	
No GSR	**With GSR**
LPF-FD Φ_F (mm)	
$\Phi_F = 0.003745 \cdot \exp\left(\frac{1.559}{\mathrm{N_{R_H}}} + \frac{264.4}{N_{V_H} - 3 - n_x}\right) + 0.02748$	$\Phi_F = 1.728 \cdot \exp\left(\frac{2.312}{\mathrm{N_{R_H}}} + \frac{294.3}{N_{V_H} - 3 - n_x}\right) + 10.07$
GEV-DV d_G (dimensionless)	
$d_G = 0.1536 \cdot \exp\left(\frac{1.559}{\mathrm{N_{R_H}}} + \frac{264.4}{N_{V_H} - 3 - n_x}\right) + 1.1272$	$d_G = 0.1567 \cdot \exp\left(\frac{2.312}{\mathrm{N_{R_H}}} + \frac{294.3}{N_{V_H} - 3 - n_x}\right) + 0.9132$

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. N_{R_H} *and* N_{V_H} *are, respectively, the number of runs per subject, and number of volumes per run, in a hypothetical dataset;* n_x *is the number of variable controlled for when calculating RSFC correlations (here, 28 in analyses without GSR, and 30 in analyses with GSR). The curve fits used to produce equations for optimal combined censoring parameters as a function of dataset size are shown in **Figure 7C-D**; cross-sections of the resulting curves are shown in **Figure 7E-H**. The optimal Φ_F/d_G ratio is demonstrated visually in **Figure S17C**. All results are from analyses using the HCP 500 dataset (n = 501; Dataset 1).*

Next, we determined optimal combined thresholds for LPF-FD and GEV-DV censoring when used together (logical OR) by seeking the global minimum of *ΔMSE-RSFC* in the space produced by the free parameter for each of the two methods (see **Table 2**). 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. 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 Parameters for Other Datasets

To allow investigators to approximate optimal volume censoring parameters for virtually any study protocol (i.e., a given number of runs per subject, and volumes per run) without having to perform the computationally intensive optimization procedures employed here, we sought to generalize our optimized LPD-FD and GEV-DV thresholds beyond the HCP 500 dataset. We first decomposed the observed between-subjects variance (across the full range of censoring thresholds) using a three-level hierarchical model (*46-48*), using combined LPF-FD and GEV-DV volume censoring (see **Figure S24**). We then estimated the between-subjects variance (**Figure S24B**) that would be observed in a dataset of a different size, but with an approximately equivalent distribution and character of motion (and pseudomotion) artifact throughout the sample (i.e., under the assumption that volumes and runs will be removed in identical proportion to the HCP 500 dataset, with proportional effects on bias and variance; see *Estimation of Optimal Censoring Thresholds for Other Datasets* in *Methods*). This allowed for ΔMSE-RSFC estimates for a dataset of any hypothetical size to be obtained across all volume censoring parameter values, which can then be minimized to obtain estimates of optimal parameter values for a hypothetical dataset.

Using the dataset-size-adjusted ΔMSE-RSFC estimates, we determined the optimal censoring thresholds for combined LPF-FD and GEV-DV censoring across a range of volumes per run and runs per subject, as shown in **Figure 8A-B**. We then used surface fitting to capture the relationship between dataset size and optimal threshold, both without GSR (adjusted R^2 = 0.9933) and with GSR (adjusted R^2

= 0.9989), as shown in **Figure 8C-D** (optimal fit parameters are shown in **Table S6**). Cross-sections of the functions for LPF-FD and GEV-DV thresholds derived from these curve fits are shown for hypothetical

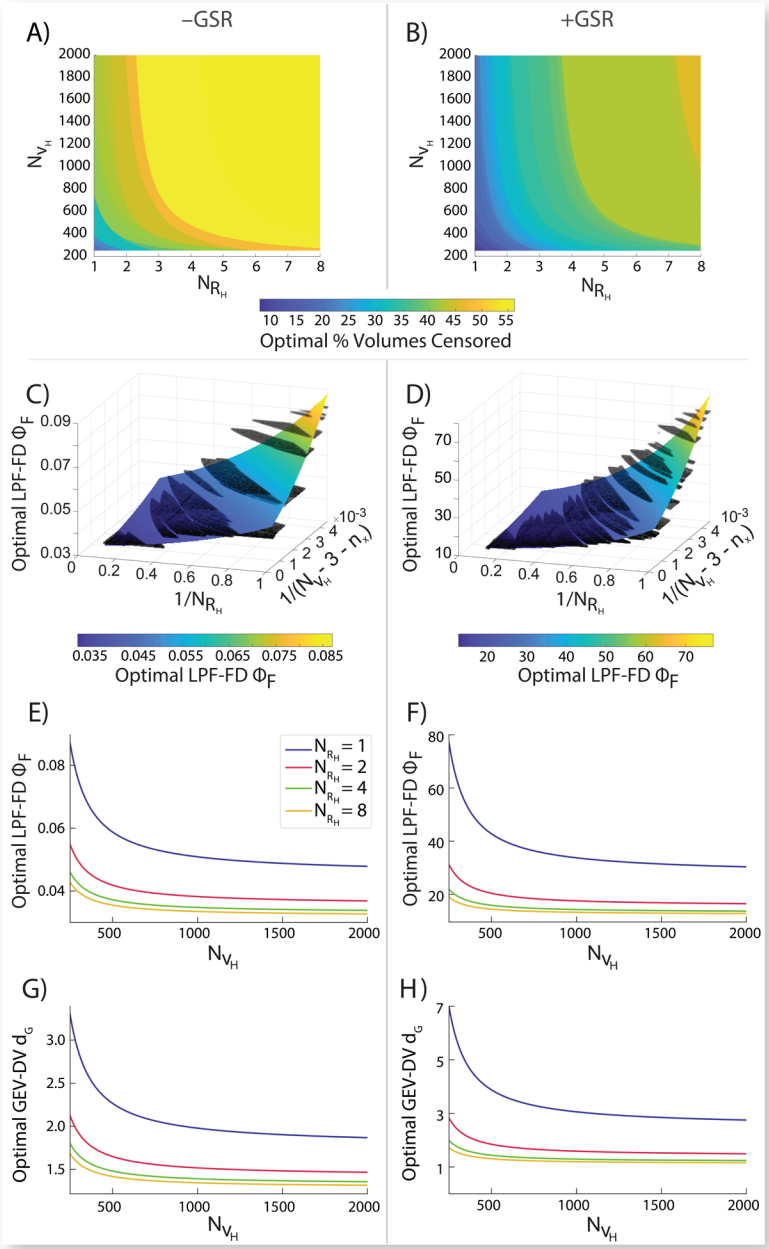

Figure 8. Optimal LPF-FD and GEV-DV volume censoring parameters as a function of acquisition protocol, determined by adjusting empirically observed estimates of ΔMSE-RSFC (see *Methods*) in the HCP 500 dataset (n = 501; *Dataset 1*). Results are shown for LPF-FD + GEV-DV censoring both without global signal regression (GSR; –GSR, left) and with GSR (+GSR, right). Optimal percent frames removed is shown as a function of number of runs (N_{R_H}) per subject and number of volumes per run (N_{V_H}), in a hypothetical dataset **(A,B)**. Surface fit of optimal LPF-FD threshold is also shown **(C,D)** with corresponding cross-sections of each surface for hypothetical datasets comprising 1, 2, 4, and 8 runs acquired per subject (N_{R_H}) **(E,F)**. Cross-sections of functions derived for optimal volume censoring parameter d_G during LPF-FD + GEV-DV censoring both without GSR and with GSR are also shown for N_{R_H} = 1, 2, 4, and 8 **(G,H).** The number of nuisance regressors (n_x) is 28 in analyses without GSR and 30 in those with GSR.

datasets of 1, 2, 4, and 8 runs (N_{R_H}) as functions N_{V_H} in **Figure 7E-H** and contour plots for the full surfaces in **Figure S17A-D**. **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}; we also supply publicly-available code implementing these formulae in MATLAB (MathWorks, Natick, MA; see *Code Availability*).

Conclusions

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), as well as in evaluating the success or failure of denoising pipelines. Specifically, we evaluated a number of dataset-quality metrics (DQMs) that have been widely used in the literature, and show that DQMs that rely on QC-FC measures, such as the correlation between RSFC correlations and subject-level QC measures (SQMs) systematically exhibit erratic behavior when assessed over a comprehensive range of censoring thresholds, raising serious questions as to their utility. Expanding on this observation, we further show that at least one "third-variable", family history of psychiatric or neurological disorder, critically confounds these DQMs as it leads to both true differences in RSFC and higher estimates of motion. In conclusion, this work argues strongly that DQMs, even those that have been widely used in the literature so far, that depend on the relationship between RSFC and SQMs such as mFD should not be used to assess motion denoising pipelines.

Next, we developed two new DQMs, MAC-RSFC and ΔMSE-RSFC, in order to serve in place of standard DQMs. The first simply quantifies the average change in RSFC values as a result of denoising, over-and-above random (i.e., not motion targeted) removal of volumes. The second is designed explicitly for optimization of denoising, in that it attempts to minimizes the error in estimates of true RSFC correlation values in a sample. Finally, we use this latter DQM to estimate optimal censoring parameters for the HCP 500 dataset, and then analytically extend those results to the full universe of possible datasets in order to provide general recommendations that should approximate optimal censoring thresholds for any dataset. We expect these methods to considerably improve the reliability and reproducibility of resting-state fMRI studies by facilitating improved data quality through motion denoising,

as well as by allowing investigators to determine optimal volume censoring methods prior to analysis, rather than during analysis as an experimenter-tunable free parameter, which will be especially valuable for pre-registration of study methods (*38*).

Limitations and Further Considerations

We wish to highlight that the novel methods presented here were specifically designed for multiband data, and would advise against generalizing these findings to any singleband, 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 in healthy young adults, 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. Datasets acquired from populations with different motion characteristics, such as children, older adults, and psychiatric or neurological patient samples, may also exhibit different optimal censoring thresholds because of differences in the prevalence of motion in a dataset relative to healthy young adults. Further, while throughout this manuscript we present results from analyses with and without GSR side-by-side in order to maximize generalizability, this procedure for optimizing censoring parameters should be repeated for denoising pipelines that utilize volume censoring together with other methods, including ICA-FIX (*25*), ICA-AROMA (*37*), temporal ICA (*49*), and DiCER (*50*), before these results can be extended to those contexts.

In addition, it is critical to note that investigators should not assume that applying the formulae 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 quite rapidly as data is removed. Rather, **Figure 8** 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 and variance. 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 may consequently leave substantial motion artifact in the dataset.

Finally, we show in **Figure S25** that 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 employ a larger number of shorter runs rather than a smaller number of long runs. Specifically, **Figure S25** strongly suggests that employing twice as many runs at half the length employed in the HCP study would have potentially resulted in substantially less participant motion than was observed.

Materials and Methods

Resting State Functional Connectivity Datasets

We employed two subsets of minimally preprocessed data (*51*) from the Human Connectome Project (HCP), denoted *Dataset 1* and *Dataset 2* and specified in *Results and Discussion* as they are used. All 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, 52*). All analyses were conducted using custom MATLAB functions. RSFC processing was similar to previously published methods (*5*), with minor changes to reduce computational and disk utilization demands.

Dataset 1 (n = 501) comprises the HCP 500 Subjects data release, with 4 runs of resting-state fMRI data per subject (3 subjects had 3 runs, and 17 had 2 runs). Each run is composed of a maximum of 1200 volumes (14 minutes 24 seconds) of data (mean = 1174.12 volumes; median = 1200 volumes), collected over two sessions, and described in detail elsewhere (*13, 35, 52*). We also produced two subsets of *Dataset 1***,** denoted *Dataset 1a* and *Dataset 1b*. In *Dataset 1a* (n = 475), we removed high-motion subjects who were dropped from analyses prior to reaching 50% of volumes censored using LPF-FD censoring, to control for the effect of subject removal when evaluating DQMs. In *Dataset 1b* (n = 315),

we removed subjects with a family history of neurological or psychiatric disease (FH+). We also produced 1,000 Monte Carlo resamplings of *Dataset 1b* in which healthy control (FH-) subjects matched to FH+ subjects based on median FD values were removed in place of FH+ subjects (n = 315; see *Comparison of High- to Low-Motion Participants After Censoring*).

Dataset 2 (n = 403) is a subset of the HCP 1200 Subjects release, comprising only subjects with 4 full (1200 volume) runs of resting-state fMRI data, with no runs flagged with QC issue A (anatomical anomalies), B (segmentation and surface QC), and C (some data acquired during periods of head coil instability), and only the first sibling from each family (based on numerical subject identifier), to eliminate non-independence due to familial relationships. This dataset includes 161 FH+ subjects and 242 FH- subjects. *Dataset 2* was used to produce three subsets consisting of 161 FH+ and 161 FH- (in each, total n = 322), selected by optimally matching FH- to FH+ subjects from ECDFs of their MPs alone (*Dataset 2a*), from the ECDFs of their both their MPs and DV values (*Dataset 2b)*, and from the means of each of their MP traces (*Dataset 2c*), as described in *Subject Matching Algorithms*.

Resting State Functional Connectivity Pre- and Post-Processing

Spheres (10 mm diameter) were drawn around 264 center coordinates reported elsewhere (*44*), and fMRI timeseries were averaged across voxels in each sphere to generate 264 timeseries for each run. Global signal (GS) was calculated as the mean signal across in-brain voxels, determined from the brainmask_fs.2.nii.gz files for each subject. 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 over participants indicated these volumes had not reached steady-state. Next, timeseries were mode 1000 normalized (multiplied by 1000 and divided by the mode all in-brain voxels), demeaned, and detrended. 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 due to filter effects at the edges of the timeseries.

Resting-State Functional Connectivity Analysis Methods

RSFC correlations were calculated as the partial correlations between each pairwise ROI, controlling for various nuisance parameters. Nuisance parameters included band-pass filtered motion parameters (MPs) and their squares (using the same filter as was applied to the RSFC timeseries), 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 (including 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 analyses, and are reported in this form throughout this manuscript. In all cases, (Z-transformed) RSFC correlations for each subject were calculated separately over runs and averaged together.

Standard Volume Censoring Methods

For standard (FD and DV) volume censoring, volumes were identified for removal when their FD or DV exceeded a threshold set for each metric separately. 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, 34, 53*). Calculation of standard DV values was carried out following volume smoothing with a 4 mm FWHM gaussian kernel in SPM 12, to produce values closer to those reported for prior datasets (*4, 5*), as unsmoothed data produces much higher values of DV; smoothed data were not used for any other purpose. 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. Subjects 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). Like standard FD and DV censoring, the aggressiveness of LPF-FD and LPF-DV was set by selecting threshold LPD-FD and LPF-DV values, denoted Φ_F and Φ_D respectively.

Adaptive LPF-DV censoring thresholds were set by maximum likelihood fitting of a generalized extreme value (GEV) distribution (*54*) to the LPF-DV values within each run, and setting the threshold LPF-DV value Φ_D separately within each run such that the ECDF 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. When LPF-FD and GEV-DV 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., logical OR).

Calculation of Subject QC and Dataset-QC Metrics (SQMs & DQMs)

QC-FC (Quality Control – Functional Connectivity) for ROI pair *k* is defined as the Pearson product-moment correlation between Z-transformed ROI pair correlations and a summary measure of subject in-scanner motion (mean FD):

$$QC\text{-}FC_{Z,mFD_k} = corr_{Pearson}\left(Z_{i,k}, mFD_i\right), \quad (1)$$

where $Z_{i,j,k}$ is the RSFC correlation for subject *i* and ROI pair *k* (averaged across all runs), and mFD_i is the mean FD value for subject *i,* averaged across all available runs for that subject. Median absolute QC-FC, shown in **Figure 1A** (for *Dataset 1a*), is defined as the median (over ROI pairs) absolute value of all QC-FC correlations, as defined by Ciric and associates (*2*):

$$Median\ absolute\ QC\text{-}FC = median\left(\left|QC\text{-}FC_{Z,mFD}\right|\right). \quad (2)$$

QC-FC null rejection rate, shown in **Figure 1B**, is defined as the proportion of all QC-FC correlations that are statistically significant (see *2, 39*), as calculated in **Equation 1**, after false discovery rate (FDR) correction for multiple comparisons (*39*). QC-FC Distance Correlation (Spearman), shown in **Figure 1C**, was calculated as the Spearman's rank-order correlation coefficient between QC-FC calculated in **Equation 1** and the distance between ROI pairs (*2-4, 7, 36*):

$$QC\text{-}FC\ Distance\ Correlation = corr_{Spearman}\left(QC\text{-}FC_{Z,mFD}, dist\right), \quad (3)$$

where $dist$ is the Euclidean distance between ROIs comprising ROI pair *k*.

To calculate High-Low Null Rejection Rate as shown in **Figure 1D**, participants were first separated into terciles based on mean FD with upper and lower terciles representing high- and low-motion participants, respectively (*5, 7, 9, 24, 37*). A Welch's (unequal variances) two-sample t-test was performed between high- and low-motion participants for each ROI pair *k* and corrected for multiple comparisons with significance threshold α = 0.000432, equivalent to 300 times the Bonferroni-corrected significance threshold for 34,716 ROI pairs derived from 264 ROIs, following Burgess and associates (*24*).

Burgess and associates also employed a two-level generalized linear model (GLM) to quantify mean and distance-dependent changes in RSFC and QC-FC correlations resulting from denoising (*24*). Here, we adapted these methods to quantify the mean and distance-dependence of RSFC correlations (rather than changes therein due to censoring), as well as the relationship of each with mFD, in *Dataset 1a*. Both the first and second level GLMs are simple linear regressions of the form, $Y = \beta_0 + \beta_1 X$, where Y is a specified dependent variable, X is a specified predictor, and β_0 and β_1 are the regression coefficients. All predictors are mean-centered before coefficients are estimated, so that β_0 estimates the mean of *Y*. Coefficients were estimated using first-level GLMs for each subject, where Y is that subject's RSFC correlations, and X is the distance between ROI pairs. In this framework, each β_0 is each subject's mean ROI pair correlation, and β_1 is the distance-dependence component of these ROI pair correlations.

We then used second-level GLMs to determine the relationship between these coefficients and a summary measure of subject motion (i.e., mean FD). Both models use (mean-centered) mFD as the *X*

variable, and then each model uses either β_0 or β_1 from the first-level model as its *Y* variable. Thus, in the model using β_0, the intercept parameter is simply the mean RSFC correlation across all subjects and ROIs **(Figure 1E)**, while the slope parameter is a measure of QC-FC (i.e., the slope of the relationship between each subject's mean RSFC and their mFD; **Figure 1G**). In contrast, in the model using β_1 from the first-level model, the intercept parameter indicates the average within-subject distance-dependence **(Figure 1H)**, while the slope parameter indicates the association of within-subject distance-dependence with subject motion (mFD; **Figure 1F**). These second-level GLMs were also repeated using the percent of volumes censored (mean centered; PVC) in place of mean FD, as shown in **Figure S7**.

Where possible, we identified optimal censoring thresholds as determined by each of these DQMs in the range of 0% to 70% volumes censored (total across the dataset), in *Dataset 1a*. We defined these optima as the value which produces the minimum magnitude of the DQM (i.e., closest to zero), or selected the first zero-crossing if multiple optima were possible in that range, for: median absolute QC-FC, QC-FC null rejection rate, QC-FC distance correlation (calculated via Spearman's rank-order correlation), high-low null rejection rate, mean QC-FC (via GLM), QC-FC slope (via GLM), mean RSFC (via GLM), and the distance-dependence of RSFC correlations (via GLM). We also noted, for variance and ΔMSE-RSFC, the point at which censoring achieves the greatest magnitude reduction (i.e., most negative change) in these statistics.

Comparison of High- to Low-Motion Participants After Censoring (Monte Carlo Simulation)

We determined the number of significant differences in RSFC correlations between high- and low-motion participants, following prior work (*3, 5, 7, 9, 37*), by splitting the sample into terciles based on median FD for each participant. We also 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 FH+ and 32 SU+ participants, respectively, to produce *Dataset 1b*, and repeated this process.

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 subjects who were FH- and SU- but who still had approximately the same mean motion as FH+ and SU+ subjects. 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 FH- and SU- participants. If this participant's median FD value was greater than the mean of the 186 excluded FH+ and SU+ 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 10^{-5}. 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.

Comparison of High- and Low-Motion Subjects Before Censoring (Contingency Table Analysis) and Evaluation of Confounds Using Permutation Tests

As above, we split subjects into terciles, here based on mean FD, and determined the number of significantly different RSFC correlations between the top and bottom terciles using a Welch's (unequal variances) two-sample t-test, using *Dataset 2* (n = 403). These 403 subjects comprise 161 FH+ and 242 FH- individuals. The 161 FH+ individuals were each optimally matched to an FH- subject on motion characteristics (without replacement), minimizing the difference between pairs of FH+ and FH- subjects

using either: 1) the Cramér–von Mises criterion computed using MPs alone (*Dataset 2a*), 2) the Cramér–von Mises criterion computed using MPs, in combination with DV (*Dataset 2b*), or 3) mean MPs (*Dataset 2c*); see *Subject Matching Algorithms.* This produced four sets of 161 FH- individuals, each resulting from matching FH- subjects to FH+ counterparts without replacement. For each of these datasets we calculated the number of significantly different RSFC correlations between high- and low-motion terciles after removing the 161 FH+ subjects, and again after instead removing the 161 FH- subjects matched by motion. We then evaluated the significance of the difference in the number of significant high-low differences obtained after removing each subset using Fisher's exact test (**Figure 3**).

We then calculated five DQMs as described in *Calculation of Subject QC and Dataset-QC Metrics (SQMs & DQMs)*, with the addition of a dichotomous nuisance regressor controlling for FH+ status, as shown in **Table 1**. Here, QC-FC Slope (Spearman's ρ), QC-FC Null Rejection Rate, and Median Absolute QC-FC were calculated normally, with the exception that QC-FC correlations were calculated as a partial correlation between ROI pair correlations and mean FD (mean centered), while controlling for FH+ status (also mean centered). Similarly, QC-FC Slope and Mean QC-FC were calculated using the two-level GLM-based approach developed by Burgess and associates (*24*), with the addition of a variable accounting for FH+ status (mean centered) and an interaction term between FH+ status and mean FD. Null expectations were generated for all five DQMs using 10,000 random permutations of FH+ membership, and p-values were obtained for the difference between the observed value of each statistic and the median of the 10,000 permutations.

Subject Matching Algorithms

Our subject matching algorithm began by dividing subjects into one subset with a positive family history of psychiatric or neurological disease (FH+), and another subset with a negative family history of psychiatric or neurological disease (FH-), as explained above. The intent was to match each subject from the FH+ group an FH- subject without replacement, such that the differences in motion metrics were minimized across the sample.

In summary, we aimed to generate a single value to quantify the similarity between the motion characteristics of any possible subject pairing, herein referred to as a subject-level motion pairing score

(S-MPS). Greater S-MPS values indicate greater difference in motion characteristics between a pair of subjects. As each subject's data comprises multiple runs, generating S-MPS values requires first generating run-level motion pairing scores (R-MPS) for each possible pairing of runs between any two subjects. With 4 runs there are 16 possible pairings for each subject pair, each with a corresponding R-MPS. We show results from four separate algorithms for producing R-MPS, which are detailed below.

The resulting 16 R-MPS values for each subject pairing can be represented as a 4x4 matrix, from which we aimed to select a single set of four run pairings (using each row and column from the 4x4 matrix only once each) that minimizes the sum of R-MPS scores. Using the MATLAB function *matchpairs* (an implementation of the Duff-Koster algorithm to solve the linear assignment problem (*55*)) we derived a single S-MPS from each 4x4 R-MPS matrix. The four R-MPS from the optimal run pairings are then summed to produce the S-MPS for that subject pair.

The resultant S-MPS for each subject pair are entered into a ranking matrix, which contains a row for each FH+ subject and a column for each FH- subject. The ranking matrix contains a single S-MPS for every possible pairing of FH+ subject to an FH- subject, such that the value in row *i* and column *j* is the S-MPS quantifying the difference in motion characteristics between FH+ subject *i* and FH- subject *j*. Next, we applied *matchpairs* to the ranking matrix to obtain the optimal subject-to-subject pairings, where optimality is defined as the minimization of the total sum of S-MPS, across all FH+ subjects without replacement.

Motion-matching was performed with three methods used to calculate run-level motion pairing scores (R-MPS). The first method (used to produce *Dataset 2a*), used in analyses shown in **Figure 3** and **Table 1**, involves the calculation of the two-sample Cramér–von Mises criterion (*43*) for each run pairing. The Cramér–von Mises criterion is a score computed via the integration of the squared difference of two empirical cumulative distribution functions (ECDFs). We calculated ECDFs for the magnitude of the derivatives of each motion parameter. The Cramér–von Mises criterion was then calculated once for each motion parameter, on the ECDFs corresponding to that motion parameter for each run. The R-MPS between those runs was calculated as the average of these criteria.

The second method (used to produce *Dataset 2b*) is identical to the first, but additionally involves the calculation of the Cramér–von Mises criterion for the magnitude of the DV vectors for each run

pairing. Both the averaged MP R-MPS and the DV R-MPS were then individually z-score normalized (over the distribution of all possible subject pairings) to control for magnitude differences, weighted evenly, and averaged to generate the S-MPS which was stored in the rankings matrix. The third and final method (used to produce *Dataset 2c*) compares runs using their motion parameters (MP). The absolute value of the derivatives of MPs (by backwards differences) was averaged for each of the two runs being compared. We calculated the magnitude (absolute value) of the difference in the two resulting means for each of the six MPs, which was then averaged over all six MPs to calculate a single R-MPS for that run pairing.

Evaluation of Volume Censoring Performance

Changes in RSFC correlations due to volume censoring were evaluated for each volume censoring method (i.e., FD, LPF-FD, DV, LPF-DV, and GEV-DV censoring), 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, 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 in *Dataset 1* as follows. For FD and LPF-FD (all units in mm): 1001 points from 0 through 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).

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*

$$\Delta Z_{i,k} = \sum_{j=1}^{N_{R_i}} \frac{Z_{i,j,k} - \frac{\sum_{p=1}^{N_{RAND}} Z_{RAND\,i,j,k,p}}{N_{RAND}}}{N_{R_i}}, \tag{4}$$

where $Z_{i,j,k}$ is the RSFC correlation (Pearson's product moment 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

$$\text{MAC-RSFC} = \sum_{k=1}^{N_{PAIR}} \frac{\sum_{i=1}^{N_S} \frac{|\Delta Z_{i,k}|}{N_S}}{N_{PAIR}}, \tag{5}$$

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 (*56*). 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). *ΔMSE-RSFC* was calculated in each ROI pair *k* and averaged over all ROI pairs, i.e.,

$$\Delta MSE\text{-}RSFC = \frac{\sum_{k=1}^{N_{PAIR}} \Delta \text{MSE-RSFC}_k}{N_{PAIR}}, \tag{6}$$

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

$$\Delta MSE\text{-}RSFC_k = \Delta\left(\frac{MSE(\hat{\theta}_k)}{N_S}\right) = \frac{MSE(\hat{\theta}_{C_k})}{N_S} - \frac{MSE(\hat{\theta}_{U_k})}{N_{S_U}}, \tag{7}$$

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:

$$MSE(\hat{\theta}_k) = E\left[(\hat{\theta}_k - \theta_k)^2\right] = \hat{\sigma}^2(\hat{\theta}_k) + Bias(\hat{\theta}_k, \theta_k)^2, \tag{8}$$

in which θ_k is the true population mean RSFC correlation for a single ROI pair *k* and $\hat{\theta}_k$ is its estimator, $\hat{\sigma}^2(\hat{\theta}_k)$ 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 removable by volume censoring (see elsewhere (*57*) for general details on the bias-variance decomposition of *MSE*). Thus, **Equation 7** can also be written as:

$$\Delta MSE\text{-}RSFC_k = \frac{\hat{\sigma}^2(\hat{\theta}_{C_k}) + Bias(\hat{\theta}_{C_k}, \theta_k)^2}{N_S} - \frac{\hat{\sigma}^2(\hat{\theta}_{U_k}) + Bias(\hat{\theta}_{U_k}, \theta_k)^2}{N_{S_U}}. \tag{9}$$

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.,

$$\Delta Bias(\hat{\theta}_k, \theta_k) = -\left|\sum_{i=1}^{N_S} \frac{\Delta Z_{i,k}}{N_S}\right|, \tag{10}$$

where $\Delta Z_{i,k}$ is defined in **Equation 4**. The change in squared bias can then be estimated as

$$\Delta\left[Bias(\hat{\theta}_k, \theta_k)^2\right] = \left|\left[Bias(\hat{\theta}_{U_k}, \theta_k) + \Delta Bias(\hat{\theta}_k, \theta_k)\right]^2 - Bias(\hat{\theta}_{U_k}, \theta_k)^2\right|. \tag{11}$$

The artifact-induced bias remaining in the estimate for ROI pair *k* after censoring may be calculated from estimates of bias in the uncensored data and the observed change in bias due to censoring:

$$Bias\left(\hat{\theta}_{C_k}, \theta_k\right)^2 = Bias\left(\hat{\theta}_{U_k}, \theta_k\right)^2 + \Delta\left[Bias\left(\hat{\theta}_k, \theta_k\right)^2\right]. \quad (12)$$

As noted above, $Bias\left(\hat{\theta}_{U_k}, \theta_k\right)$ is the total sample-wide motion-induced bias in RSFC correlations that is removable 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, $\Delta Bias\left(\hat{\theta}_k, \theta_k\right)$, 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 was done 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 11** 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 S26.**

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

Optimal parameters for combined LPF-FD and GEV-DV censoring were determined with an optimization procedure minimizing *ΔMSE-RSFC* in *Dataset 1*. We used a particle swarm global optimization algorithm (*particleswarm* in MATLAB; *58-60*), followed by simulated annealing (*simulannealbnd*; *61-63*), each followed by a pattern search local optimization algorithm (*patternsearch* with default settings*; 64-69*) to refine results. Particle swarm optimization used 44 particles, 1000 stall iterations and an initial neighborhood size of 1. 40 particles 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). Simulated annealing used *particleswarm* optima (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 length temperature and direction chosen uniformly at random (i.e., *'annealingfast'*).

Estimation of Optimal Censoring Thresholds for Other Datasets

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). Here, we define this hypothetical dataset to have N_{V_H} volumes per run and N_{R_H} runs per subject.

We first consider a variance decomposition of a given RSFC correlation for ROI pair *k*, based on a three-level hierarchical model with unweighted means (*46-48*). At the top level, the observed between-subjects variance can be described using the following relationship:

$$\hat{\sigma}^2(\hat{\theta}_k) = \hat{\tau}_S^2(\hat{\theta}_k) + \frac{\hat{\sigma}_R^2(\hat{\theta}_k)}{\widetilde{N}_R} \ , \tag{13}$$

where $\hat{\sigma}^2(\hat{\theta}_k)$ is the observed between-subjects variance for ROI pair *k*, $\hat{\sigma}_R^2(\hat{\theta}_k)$ is the observed between-runs variance for ROI pair *k*, $\hat{\tau}_S^2(\hat{\theta}_k)$ is the estimated true between-subjects variance for ROI pair k, and $\widetilde{N}_R$ is the harmonic mean of the number of runs across all subjects in the study.

The known variance of a Z-transformed Pearson's product-moment correlation for subject *i* and run *j* is*:*

$$\sigma_Z^2(\hat{\theta}_{i,j,k}) = \frac{1}{N_{V_{i,j}} - 3 - n_x} \ , \tag{14}$$

where $N_{V_{i,j}}$ is the number of volumes remaining in the analysis after volume censoring for subject *i* and run *j*, and n_x is the number of variables controlled for when calculating RSFC partial correlations (here, n_x = 28 in analyses without GSR; n_x = 30 in analyses with GSR). Note that this formula is independent of *k*, as an equivalent number of volumes is censored for all ROI pairs.

For each subject *i*, we obtain $\sigma_Z^2(\hat{\theta}_{i,k})$ as the mean of $\sigma_Z^2(\hat{\theta}_{i,j,k})$ across all runs *j*, and can thus write the observed between-runs variance as:

$$\hat{\sigma}_R^2(\hat{\theta}_{i,k}) = \hat{\tau}_R^2(\hat{\theta}_{i,k}) + \sigma_Z^2(\hat{\theta}_{i,k}) \ , \tag{15}$$

where, for subject *i* and ROI pair *k*, $\hat{\tau}_R^2(\hat{\theta}_{i,k})$ is the estimated true between-runs variance, and $\sigma_Z^2(\hat{\theta}_{i,k})$ is the total contribution of sampling error to the run-level estimate of the RSFC correlation for ROI pair *k*. As

$\hat{\sigma}_R^2(\hat{\theta}_{i,k})$ can be calculated directly (it is the variance over runs within a subject) and $\sigma_Z^2(\hat{\theta}_{i,k})$ is known analytically (see **Equation 14**), the true between-runs variance for each subject can be estimated:

$$\hat{\tau}_R^2(\hat{\theta}_{i,k}) = \hat{\sigma}_R^2(\hat{\theta}_{i,k}) - \sigma_Z^2(\hat{\theta}_{i,k}), \tag{16}$$

thus providing an estimate of $\hat{\tau}_R^2(\hat{\theta}_{i,k})$. We further calculate $\hat{\sigma}_R^2(\hat{\theta}_k)$ as the mean of $\hat{\sigma}_R^2(\hat{\theta}_{i,k})$ across all subjects *i*, allowing us to obtain estimates of the true between-subjects variance:

$$\hat{\tau}_S^2(\hat{\theta}_k) = \hat{\sigma}^2(\hat{\theta}_k) - \frac{\hat{\sigma}_R^2(\hat{\theta}_k)}{\widetilde{N}_R}. \tag{17}$$

We can alternatively write **Equation 13** as follows, after defining $\hat{\tau}_R^2(\hat{\theta}_k)$ as the across-subjects mean of all $\hat{\tau}_R^2(\hat{\theta}_{i,k})$ and $\sigma_Z^2(\hat{\theta}_k)$ as the across-subjects mean of all $\sigma_Z^2(\hat{\theta}_{i,k})$:

$$\hat{\sigma}^2(\hat{\theta}_k) = \hat{\tau}_S^2(\hat{\theta}_k) + \frac{\hat{\tau}_R^2(\hat{\theta}_k) + \sigma_Z^2(\hat{\theta}_k)}{\widetilde{N}_R}. \tag{18}$$

Using this relationship, we can adjust the estimates of between-subjects variance observed in the HCP500 dataset before censoring, $\hat{\sigma}^2(\hat{\theta}_{U_k})$, and that observed after censoring, $\hat{\sigma}^2(\hat{\theta}_{C_k})$, to instead reflect a dataset with a hypothetical number of runs per subject, and volumes per run. We assume that in such 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 (*Dataset 1*).

First, we define $\chi = \frac{\widetilde{N}_{R_C}}{\widetilde{N}_{R_U}}$, where $\widetilde{N}_{R_U}$ and $\widetilde{N}_{R_C}$ are the across-subjects harmonic mean of the number of runs per subject before and after censoring, respectively, in the HCP500 dataset. We further define for each run *j* of subject *i*, $\gamma_{i,j} = \frac{N_{V_{C_{i,j}}}}{N_{V_{U_{i,j}}}}$, where $N_{V_{U_{i,j}}}$ and $N_{V_{C_{i,j}}}$ are respectively the number of volumes in run *j* of subject *i* before and after volume censoring. Next, we calculate the observed sampling variance before censoring, $\sigma_Z^2\left(\hat{\theta}_{U_{i,j,k}}\right)$, from **Equation 14**:

$$\sigma_Z^2\left(\hat{\theta}_{U_{i,j,k}}\right) = \frac{1}{N_{V_{i,j}} - 3 - n_x}. \tag{19}$$

Similarly, we calculate the observed sampling variance after censoring, $\sigma_Z^2\left(\hat{\theta}_{C_{i,k}}\right)$:

$$\sigma_Z^2\left(\hat{\theta}_{C_{i,j,k}}\right) = \frac{1}{(\gamma_{i,j} \cdot N_{V_{i,j}}) - 3 - n_x} . \tag{20}$$

We can follow this framework to obtain a hypothetical sampling variance before censoring, $\sigma_{Z_H}^2\left(\hat{\theta}_{U_{i,j,k}}\right)$, for each run to reflect a hypothetical dataset with uniformly N_{V_H} volumes per run as follows:

$$\sigma_{Z_H}^2\left(\hat{\theta}_{U_{i,j,k}}\right) = \frac{1}{N_{V_H} - 3 - n_x} . \tag{21}$$

Next, we obtain the hypothetical sampling variance after censoring, $\sigma_{Z_H}^2\left(\hat{\theta}_{C_{i,j,k}}\right)$:

$$\hat{\sigma}_{Z_H}^2\left(\hat{\theta}_{C_{i,j,k}}\right) = \frac{1}{(\gamma_{i,j} \cdot N_{V_H}) - 3 - n_x} . \tag{22}$$

We can then calculate the mean hypothetical sampling variance for each subject, $\hat{\sigma}_{Z_H}^2\left(\hat{\theta}_{C_{i,k}}\right)$, as the algebraic mean of $\hat{\sigma}_{Z_H}^2\left(\hat{\theta}_{C_{i,j,k}}\right)$ across runs, and the hypothetical sampling variance across all subjects, $\hat{\sigma}_{Z_H}^2\left(\hat{\theta}_{C_k}\right)$, as the mean of all $\hat{\sigma}_{Z_H}^2\left(\hat{\theta}_{C_{i,k}}\right)$. This also holds for $\sigma_Z^2\left(\hat{\theta}_{U_k}\right)$ and $\sigma_Z^2\left(\hat{\theta}_{C_k}\right)$, which are respectively the mean first across runs, and then across subjects, of $\sigma_{Z_H}^2\left(\hat{\theta}_{U_{i,j,k}}\right)$ and $\hat{\sigma}_{Z_H}^2\left(\hat{\theta}_{C_{i,j,k}}\right)$.

From this, we can adjust **Equation 18** to reflect a hypothetical dataset as follows in uncensored data:

$$\hat{\sigma}_H^2\left(\hat{\theta}_{U_k}\right) = \hat{\tau}_S^2\left(\hat{\theta}_{U_k}\right) - \frac{\hat{\tau}_R^2\left(\hat{\theta}_{U_k}\right) + \sigma_Z^2\left(\hat{\theta}_{U_k}\right)}{\widetilde{N}_{R_U}} + \frac{\hat{\tau}_R^2\left(\hat{\theta}_{U_k}\right) + \sigma_{Z_H}^2\left(\hat{\theta}_{U_k}\right)}{N_{V_H}}, \tag{23}$$

where $\widetilde{N}_{R_U}$ is the (across-subjects) harmonic mean of the number of runs in the (actual) HCP500 dataset (*Dataset 1*) before censoring. We can also calculate the hypothetical observed between-subjects variance in censored data:

$$\hat{\sigma}_H^2\left(\hat{\theta}_{C_k}\right) = \hat{\tau}_S^2\left(\hat{\theta}_{C_k}\right) - \frac{\hat{\tau}_R^2\left(\hat{\theta}_{C_k}\right) + \sigma_Z^2\left(\hat{\theta}_{C_k}\right)}{\widetilde{N}_{R_C}} + \frac{\hat{\tau}_R^2\left(\hat{\theta}_{C_k}\right) + \sigma_{Z_H}^2\left(\hat{\theta}_{C_k}\right)}{(\chi \cdot N_{V_H})} , \tag{24}$$

where $\widetilde{N}_{R_C}$ is the (across-subjects) harmonic mean of the number of runs in the HCP500 dataset after censoring.

This can then be used to obtain an estimate of *ΔMSE-RSFC* for this hypothetical dataset by replacing $\hat{\sigma}^2\left(\hat{\theta}_{C_k}\right)$ terms in **Equation 9** with $\hat{\sigma}_H^2\left(\hat{\theta}_{C_k}\right)$ and, assuming an equal proportion of subjects removed due to volume censoring in this hypothetical dataset as in the HCP500 dataset,

$$\Delta MSE\text{-}RSFC_{H_k} = \frac{\hat{\sigma}_H^2\left(\hat{\theta}_{C_k}\right) + Bias\left(\hat{\theta}_{C_k}, \theta_k\right)^2}{\zeta N_{S_H}} - \frac{\hat{\sigma}^2\left(\hat{\theta}_{U_k}\right) + Bias\left(\hat{\theta}_{U_k}, \theta_k\right)^2}{N_{S_H}}, \tag{25}$$

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 (*Dataset 1*). Finally, we can calculate the average $\Delta MSE\text{-}RSFC_{H_k}$ across all ROI pairs using **Equation 6**:

$$\Delta\text{MSE-RSFC}_H = \sum_{k=1}^{N_{PAIR}} \frac{\Delta MSE\text{-}RSFC_{H_k}}{N_{PAIR}} . \tag{26}$$

Using **Equations 21-24**, we adjusted variance estimates from the HCP500 dataset (*Dataset 1*) 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 0.01, and N_{V_H} was varied in the range [250, 2000], in steps of 1 from 100 through 500, and in steps of 5 from 505 through 2000. ΔMSE-RSFC was recalculated using **Equations 25-26** by adjusting estimates for variance for each combination of N_{R_H} and N_{V_H}, and across volume censoring parameters. Volume censoring parameters were varied using the optimal ratio of LPF-FD Φ_F to GEV-DV d_G when used in tandem as shown in **Table 2** (i.e., along the vectors shown in **Figure S17E**), separately for analyses with GSR and without GSR. A total of 7,522 sets of volume censoring parameters were sampled, including 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 (e.g**., Figure 8A-B**) are displayed after smoothing using a 2D Gaussian smoothing kernel with a standard deviation of 1. Surface fits for Φ_F as a function of hypothetical dataset size, shown in **Figure 8C-D**, were carried out using the Curve Fitting Toolbox in MATLAB R2018b (*cftool*), using the trust-region algorithm, coefficient starting values of [1,1,1,1], coefficient and function termination tolerances (*'TolX'* and *'TolFun'*) of 10^{-12}, minimum change in coefficient values per iteration of 10^{-12}, and a maximum change in coefficient values per iteration of 1; all fits converged.

References

1. https://www.nimh.nih.gov/about/strategic-planning-reports/strategic-research-priorities/index.shtml. (2017), vol. 2018.
2. R. Ciric, D. H. Wolf, J. D. Power, D. R. Roalf, G. L. Baum, K. Ruparel, R. T. Shinohara, M. A. Elliott, S. B. Eickhoff, C. Davatzikos, R. C. Gur, R. E. Gur, D. S. Bassett, T. D. Satterthwaite, Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* **154**, 174-187 (2017).
3. L. Parkes, B. Fulcher, M. Yucel, A. Fornito, An evaluation of the efficacy, reliability, and sensitivty of motion correction strategies for resting-state functional MRI. *NeuroImage* **171**, 415-436 (2018).
4. J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, S. E. Petersen, Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142-2154 (2012).
5. J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar, S. E. Petersen, Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320-341 (2014).
6. J. D. Power, B. L. Schlaggar, S. E. Petersen, Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage* **105**, 536-551 (2015).
7. T. D. Satterthwaite, M. A. Elliott, R. T. Gerraty, K. Ruparel, J. Loughead, M. E. Calkins, S. B. Eickhoff, H. Hakonarson, R. C. Gur, R. E. Gur, D. H. Wolf, 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. T. D. Satterthwaite, D. H. Wolf, J. Loughead, K. Ruparel, M. A. Elliott, H. Hakonarson, R. C. Gur, R. E. Gur, 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. K. R. Van Dijk, M. R. Sabuncu, R. L. Buckner, The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* **59**, 431-438 (2012).

10. C. G. Yan, B. Cheung, C. Kelly, S. Colcombe, R. C. Craddock, A. Di Martino, Q. Li, X. N. Zuo, F. X. Castellanos, M. P. Milham, A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *Neuroimage* **76**, 183-201 (2013).

11. J. D. Power, B. M. Silver, M. R. Silverman, E. L. Ajodan, D. J. Bos, R. M. Jones, Customized head molds reduce motion during resting state fMRI scans. *Neuroimage* **189**, 141-149 (2019).

12. S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, K. Ugurbil, 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. K. Ugurbil, J. Xu, E. J. Auerbach, S. Moeller, A. T. Vu, J. M. Duarte-Carvajalino, C. Lenglet, X. Wu, S. Schmitter, P. F. Van de Moortele, J. Strupp, G. Sapiro, F. De Martino, D. Wang, N. Harel, M. Garwood, L. Chen, D. A. Feinberg, S. M. Smith, K. L. Miller, S. N. Sotiropoulos, S. Jbabdi, J. L. Andersson, T. E. Behrens, M. F. Glasser, D. C. Van Essen, E. Yacoub, W. U.-M. H. Consortium, Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage* **80**, 80-104 (2013).

14. K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. Andersson, L. Griffanti, G. Douaud, T. W. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P. M. Matthews, S. M. Smith, Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* **19**, 1523-1536 (2016).

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

16. B. J. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, D. M. Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan, C. A. Orr, T. D. Wager, M. T. Banich, N. K. Speer, M. T. Sutherland, M. C. Riedel, A. S. Dick, J. M. Bjork, K. M. Thomas, B. Chaarani, M. H. Mejia, D. J. Hagler, Jr., M. Daniela Cornejo, C. S. Sicat, M. P. Harms, N. U. F. Dosenbach, M. Rosenberg, E. Earl, H. Bartsch, R. Watts, J. R. Polimeni, J. M. Kuperman, D. A. Fair, A. M. Dale, A. I. A. Workgroup, The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cogn Neurosci* **32**, 43-54 (2018).

17. N. D. Volkow, G. F. Koob, R. T. Croyle, D. W. Bianchi, J. A. Gordon, W. J. Koroshetz, E. J. Perez-Stable, W. T. Riley, M. H. Bloch, K. Conway, B. G. Deeds, G. J. Dowling, S. Grant, K. D. Howlett, J. A. Matochik, G. D. Morgan, M. M. Murray, A. Noronha, C. Y. Spong, E. M. Wargo, K. R. Warren, S. R. B. Weiss, The conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev Cogn Neurosci* **32**, 4-7 (2018).

18. D. J. Hagler, Jr., S. Hatton, M. D. Cornejo, C. Makowski, D. A. Fair, A. S. Dick, M. T. Sutherland, B. J. Casey, D. M. Barch, M. P. Harms, R. Watts, J. M. Bjork, H. P. Garavan, L. Hilmer, C. J. Pung, C. S. Sicat, J. Kuperman, H. Bartsch, F. Xue, M. M. Heitzeg, A. R. Laird, T. T. Trinh, R. Gonzalez, S. F. Tapert, M. C. Riedel, L. M. Squeglia, L. W. Hyde, M. D. Rosenberg, E. A. Earl, K. D. Howlett, F. C. Baker, M. Soules, J. Diaz, O. R. de Leon, W. K. Thompson, M. C. Neale, M. Herting, E. R. Sowell, R. P. Alvarez, S. W. Hawes, M. Sanchez, J. Bodurka, F. J. Breslin, A. S. Morris, M. P. Paulus, W. K. Simmons, J. R. Polimeni, A. van der Kouwe, A. S. Nencka, K. M. Gray, C. Pierpaoli, J. A. Matochik, A. Noronha, W. M. Aklin, K. Conway, M. Glantz, E. Hoffman, R. Little, M. Lopez, V. Pariyadath, S. R. Weiss, D. L. Wolff-Hughes, R. DelCarmen-Wiggins, S. W. Feldstein Ewing, O. Miranda-Dominguez, B. J. Nagel, A. J. Perrone, D. T. Sturgeon, A. Goldstone, A. Pfefferbaum, K. M. Pohl, D. Prouty, K. Uban, S. Y. Bookheimer, M. Dapretto, A. Galvan, K. Bagot, J. Giedd, M. A. Infante, J. Jacobus, K. Patrick, P. D. Shilling, R. Desikan, Y. Li, L. Sugrue, M. T. Banich, N. Friedman, J. K. Hewitt, C. Hopfer, J. Sakai, J. Tanabe, L. B. Cottler, S. J. Nixon, L. Chang, C. Cloak, T. Ernst, G. Reeves, D. N. Kennedy, S. Heeringa, S. Peltier, J. Schulenberg, C. Sripada, R. A. Zucker, W. G. Iacono, M. Luciana, F. J. Calabro, D. B. Clark, D. A. Lewis, B. Luna, C. Schirda, T. Brima, J. J. Foxe, E. G. Freedman, D. W. Mruzek, M. J. Mason, R. Huber, E. McGlade, A. Prescot, P. F. Renshaw, D. A. Yurgelun-Todd, N. A. Allgaier, J. A. Dumas, M. Ivanova, A. Potter, P. Florsheim, C. Larson, K. Lisdahl, M. E. Charness, B. Fuemmeler, J. M. Hettema, H. H. Maes, J. Steinberg, A. P. Anokhin, P. Glaser, A. C. Heath, P. A. Madden, A. Baskin-Sommers, R. T. Constable, S. J. Grant, G. J. Dowling, S. A. Brown, T. L. Jernigan, A. M. Dale, Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *Neuroimage*, 116091 (2019).

19. D. A. Fair, O. Miranda-Dominguez, A. Z. Snyder, A. Perrone, E. A. Earl, A. N. Van, J. M. Koller, E. Feczko, M. D. Tisdall, A. van der Kouwe, R. L. Klein, A. E. Mirro, J. M. Hampton, B. Adeyemo, T. O. Laumann, C. Gratton, D. J. Greene, B. L. Schlaggar, D. J. Hagler, Jr., R. Watts, H. Garavan, D. M. Barch, J. T. Nigg, S. E. Petersen, A. M. Dale, S. W. Feldstein-Ewing, B. J. Nagel, N. U. F. Dosenbach, Correction of respiratory artifacts in MRI head motion estimates. *Neuroimage* **208**, 116400 (2020).

20. J. D. Power, C. J. Lynch, B. M. Silver, M. J. Dubin, A. Martin, R. M. Jones, Distinctions among real and apparent respiratory motions in human fMRI data. *Neuroimage* **201**, 116041 (2019).

21. J. Williams, J. Van Snellenberg, Artifact idnetification and removal in simultaneous multi-slice (multiband) fMRI. *Biological Psychiatry* **83**, S439-S440 (2018).

22. J. Williams, J. Van Snellenberg, Optimized volume censoring in simultaneous multi-slice (multiband) resting-state fMRI. *Biological Psychiatry* **85**, S188 (2019).

23. J. C. Williams, J. X. Van Snellenberg, Motion denoising of multiband resting state functional connectivity MRI data: An improved volume censoring method. *BioRxiv*, 49 (2019).

24. G. C. Burgess, S. Kandala, D. Nolan, T. O. Laumann, J. D. Power, B. Adeyemo, M. P. Harms, S. E. Petersen, D. M. Barch, 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).

25. L. Griffanti, G. Salimi-Khorshidi, C. F. Beckmann, E. J. Auerbach, G. Douaud, C. E. Sexton, E. Zsoldos, K. P. Ebmeier, N. Filippini, C. E. Mackay, S. Moeller, J. Xu, E. Yacoub, G. Baselli, K. Ugurbil, K. L. Miller, S. M. Smith, ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage* **95**, 232-247 (2014).

26. G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, S. M. Smith, Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90**, 449-468 (2014).

27. J. D. Power, C. J. Lynch, B. Adeyemo, S. E. Petersen, A Critical, Event-Related Appraisal of Denoising in Resting-State fMRI Studies. *Cereb Cortex*, (2020).

28. Z. S. Saad, S. J. Gotts, K. Murphy, G. Chen, H. J. Jo, A. Martin, R. W. Cox, Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. *Brain Connect* **2**, 25-32 (2012).

29. J. D. Power, M. Plitt, T. O. Laumann, A. Martin, Sources and implications of whole-brain fMRI signals in humans. *Neuroimage* **146**, 609-625 (2017).

30. K. Murphy, M. D. Fox, Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage* **154**, 169-173 (2017).

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

32. M. D. Fox, D. Zhang, A. Z. Snyder, M. E. Raichle, The global signal and observed anticorrelated resting state brain networks. *J Neurophysiol* **101**, 3270-3283 (2009).

33. T. E. Nichols, Notes on Creating a Standardized Version of DVARS. *arXiv e-prints*, arXiv:1704.01469 (2017).

34. S. Afyouni, T. E. Nichols, Insight and inference for DVARS. *Neuroimage* **172**, 291-312 (2018).

35. D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W. U.-M. H. Consortium, The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62-79 (2013).

36. J. Muschelli, M. B. Nebel, B. S. Caffo, A. D. Barber, J. J. Pekar, S. H. Mostofsky, Reduction of motion-related artifacts in resting state fMRI using aCompCor. *Neuroimage* **96**, 22-35 (2014).

37. R. H. R. Pruim, M. Mennes, J. K. Buitelaar, C. F. Beckmann, Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. *Neuroimage* **112**, 278-287 (2015).

38. K. J. Gorgolewski, R. A. Poldrack, A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research. *PLoS Biol* **14**, e1002506 (2016).

39. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289-300 (1995).

40. L. L. Zeng, D. Wang, M. D. Fox, M. Sabuncu, D. Hu, M. Ge, R. L. Buckner, H. Liu, Neurobiological basis of head motion in brain imaging. *Proc Natl Acad Sci U S A* **111**, 6058-6062 (2014).

41. J. S. Siegel, A. Mitra, T. O. Laumann, B. A. Seitzman, M. Raichle, M. Corbetta, A. Z. Snyder, Data Quality Influences Observed Links Between Functional Connectivity and Behavior. *Cereb Cortex* **27**, 4492-4502 (2017).

42. T. A. W. Bolton, V. Kebets, E. Glerean, D. Zoller, J. Li, B. T. T. Yeo, C. Caballero-Gaudes, D. Van De Ville, Agito ergo sum: Correlates of spatio-temporal motion characteristics during fMRI. *Neuroimage* **209**, 116433 (2020).

43. T. W. Anderson, On the Distribution of the Two-Sample Cramér-von Mises Criterion. *The Annals of Mathematical Statistics* **33**, 1148-1159 (1962).

44. J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, S. E. Petersen, Functional network organization of the human brain. *Neuron* **72**, 665-678 (2011).

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

46. T. A. B. Snijders, R. J. Bosker, *Multilevel analysis : an introduction to basic and advanced multilevel modeling* (Sage, Los Angeles, ed. 2nd, 2012), pp. xi, 354 p.

47. S. R. Searle, G. Casella, C. E. McCulloch, *Variance components* (Wiley series in probability and mathematical statistics Applied probability and statistics, Wiley, New York, 1992), pp. xxiii, 501 p.

48. R. Wilms, R. Lanwehr, A. Kastenmuller, Do We Overestimate the Within-Variability? The Impact of Measurement Error on Intraclass Coefficient Estimation. *Front Psychol* **11**, 825 (2020).

49. M. F. Glasser, T. S. Coalson, J. D. Bijsterbosch, S. J. Harrison, M. P. Harms, A. Anticevic, D. C. Van Essen, S. M. Smith, Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. *Neuroimage* **181**, 692-717 (2018).

50. K. M. Aquino, B. D. Fulcher, L. Parkes, K. Sabaroedin, A. Fornito, Identifying and removing widespread signal deflections from fMRI data: Rethinking the global signal regression problem. *Neuroimage* **212**, 116614 (2020).

51. M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, M. Jenkinson, W. U.-M. H. Consortium, The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105-124 (2013).

52. S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, M. Kelly, T. Laumann, K. L. Miller, S. Moeller, S. Petersen, J. Power, G. Salimi-Khorshidi, A. Z. Snyder, A. T. Vu, M. W. Woolrich, J. Xu, E. Yacoub, K. Ugurbil, D. C. Van Essen, M. F. Glasser, W. U.-M. H. Consortium, Resting-state fMRI in the Human Connectome Project. *Neuroimage* **80**, 144-168 (2013).

53. C. D. Smyser, A. Z. Snyder, J. J. Neil, Functional connectivity MRI in infants: exploration of the functional organization of the developing brain. *Neuroimage* **56**, 1437-1452 (2011).

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

55. I. Duff, J. Koster, On Algorithms For Permuting Large Entries to the Diagonal of a Sparse Matrix. *SIAM J. Matrix Anal. Appl.* **22**, 973-996 (2001).

56. B. Efron, R. Tibshirani, *An introduction to the bootstrap* (Monographs on statistics and applied probability, Chapman & Hall, New York, 1993), pp. xvi, 436 p.

57. D. D. Wackerly, W. Mendenhall, R. L. Scheaffer, *Mathematical statistics with applications* (Thomson Brooks/Cole, Belmont, CA, ed. 7th, 2008), pp. xxii, 912 p.

58. Y. Shi, R. Eberhart, in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*. (1998), pp. 69-73.

59. J. Kennedy, R. Eberhart, in *Proceedings of ICNN'95 - International Conference on Neural Networks*. (1995), vol. 4, pp. 1942-1948 vol.1944.

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

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

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

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

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

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

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

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

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

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

70. D. A. Fair, O. Miranda-Dominguez, A. Z. Snyder, A. Perrone, E. A. Earl, A. N. Van, J. M. Koller, E. Feczko, R. L. Klein, A. E. Mirro, J. M. Hampton, B. Adeyemo, T. O. Laumann, C. Gratton, D. J. Greene, B. L. Schlaggar, D. Hagler, R. Watts, H. Garavan, D. M. Barch, J. T. Nigg, S. E. Petersen, A. Dale, S. W. Feldstein-Ewing, B. J. Nagel, N. U. F. Dosenbach, Correction of respiratory artifacts in MRI head motion estimates. *bioRxiv*, (2018).

71. J. D. Power, C. J. Lynch, M. J. Dubin, B. M. Silver, A. Martin, R. M. Jones, Characteristics of respiratory measures in young adults scanned at rest, including systematic changes and "missed" deep breaths. *Neuroimage* **204**, 116234 (2019).

72. G. Chen, P. A. Taylor, S. P. Haller, K. Kircanski, J. Stoddard, D. S. Pine, E. Leibenluft, M. A. Brotman, R. W. Cox, Intraclass correlation: Improved modeling approaches and applications for neuroimaging. *Human brain mapping* **39**, 1187-1206 (2018).

73. R. M. Birn, M. D. Cornejo, E. K. Molloy, R. Patriat, T. B. Meier, G. R. Kirk, V. A. Nair, M. E. Meyerand, V. Prabhakaran, The influence of physiological noise correction on test-retest reliability of resting-state functional connectivity. *Brain Connect* **4**, 511-522 (2014).

74. P. E. Shrout, J. L. Fleiss, Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86**, 420-428 (1979).

Acknowledgements: Mark Slifstein, Hongshik Ahn, Joseph Schwartz, Yuefan Deng, and Anissa Abi-Dargham provided helpful commentary on various aspects of the methods presented here. Zu Jie Zheng, Alexander Eichert, Eilon Silver-Frankel, Hung-Wei (Bernie) Chen, and Sameera Abeykoon assisted with deploying analysis software to the SeaWulf cluster and generating figures. Mutahira Bhatti, Umaimah Nawaz, Justin Beutel, and Ayman A. Khan provided assistance with editing figures and tables presented in this manuscript. **T**he authors would additionally like to thank Stony Brook Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science (IACS) at Stony Brook University for access to the high-performance SeaWulf computing system, which was made possible by a $1.4M National Science Foundation grant (#1531492). IACS staff additionally provided critical technical assistance for the completion of this work, with notable support from Firat Coskun. **Funding:** Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award numbers K01 MH 107763 to JXVS and F30 MH 122136 to JCW. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. **Author contributions:** Conceptualization, J.C.W. and J.X.V.S.; Data Curation, J.C.W. and J.X.V.S.; Formal Analysis, J.C.W., P.N.T., and J.X.V.S.; Funding Acquisition, J.C.W. and J.X.V.S.; Investigation, J.C.W., P.N.T., and J.X.V.S; Methodology, J.C.W., P.N.T., J.R.L., and J.X.V.S.; Project Administration, J.C.W. and J.X.V.S.; Resources, J.X.V.S.; Software, J.C.W., P.N.T., J.R.L., and J.X.V.S.; Supervision, J.X.V.S.; Validation, J.C.W., P.N.T., and J.X.V.S.; Visualization, J.C.W., P.N.T., and J.X.V.S.; Writing – Original Draft, J.C.W., P.N.T., J.R.L., and J.X.V.S.; Writing – Review &

Editing, J.C.W., P.N.T., J.R.L., and J.X.V.S. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The data employed here are publicly available from the Human Connectome Project at https://db.humanconnectome.org, and include all resting state fMRI data from the *S500 Subjects Release* and a subset of the resting-state fMRI data from the *S1200 Subjects Release* (*13, 35, 51, 52*), as described in *Resting-State Functional Connectivity Dataset*. Data were provided by the Human Connectome Project (*35*), WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; U54 MH 091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. **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.

Supplementary Materials

Results and Discussion

Materials and Methods

Tables S1 – S6

Figures S1 – S29