Brain Imaging and Behavior (2015) 9:115–127
DOI 10.1007/s11682-014-9351-7

SI: DEVELOPING BRAIN

One dataset, many conclusions: BOLD variability's complicated relationships with age and motion artifacts

Benjamin O. Turner · Brian Lopez · Tyler Santander · Michael B. Miller

Published online: 9 January 2015

Abstract In recent years, the variability of the blood-oxygen level dependent (BOLD) signal has received attention as an informative measure in its own right. At the same time, there has been growing concern regarding the impact of motion in fMRI, particularly in the domain of resting state studies. Here, we demonstrate that, not only does motion (among other confounds) exert an influence on the results of a BOLD variability analysis of task-related fMRI data—but, that the exact method used to deal with this influence has at least as large an effect as the motion itself. This sensitivity to relatively minor methodological changes is particularly concerning as studies begin to take on a more applied bent, and the risk of mischaracterizing the relationship between BOLD variability and various individual difference variables (for instance, disease progression) acquires real-world relevance.

Keywords Individual differences · Correlation analysis · FMRI analysis methods · Confound correction

Introduction

As the field of fMRI research has moved beyond simply characterizing the mean activity associated with the performance of various tasks, the variability of the BOLD signal has begun to receive greater attention. For instance, a series of related studies by Grady and colleagues (Garrett et al. 2010; 2011; 2012; 2013) have investigated the variability of the BOLD timeseries, and the relationship between variability in different regions and various individual difference measures. Among the findings reported in this line of papers: that BOLD variability represents a largely orthogonal source of information from BOLD mean, and that variability shows a reliable relationship with age across a number of brain regions (Garrett et al. 2010); that overall BOLD variability seems to correspond with more optimal task performance (Garrett et al. 2011); and, that BOLD variability is linked to mental set (e.g., task vs. rest), and the degree of higher-order variability (i.e., variability of variability across conditions) is related to task performance (Garrett et al. 2012).

Other researchers have likewise begun to investigate BOLD variability as a measure of interest, both as a way to understand the brain activity underlying particular processes or tasks, and as a way to investigate changes in the brain associated with, e.g., aging or disease processes. Several such investigations have moved beyond simply examining the variance of the BOLD timeseries. For instance, He (2011) examines the degree to which the BOLD signal in different areas follows a power-law in its spectral properties, and how this relates to task or rest. Using a different method of characterizing BOLD timeseries variability—namely, approximate entropy, a measure of signal complexity—Liu et al. (2013) demonstrate that BOLD signal complexity decreases with age, and that this decrease is related to cognitive decline in a group of participants with familial Alzheimer's disease. Similarly, Samanez-Larkin et al. (2010) show that variability in the nucleus accumbens (measured by mean squared successive difference; see also Mohr and Nagel 2010) mediates the degree of age-related loss of

B. O. Turner (✉) · B. Lopez · T. Santander · M. B. Miller
Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA, USA
e-mail: turner@psych.ucsb.edu

M. B. Miller
e-mail: michael.miller@psych.ucsb.edu

optimality in making risky choices in a financial decision-making task. Using the same measure, Leo et al. (2012) report that several brain areas evince greater variability in blind individuals compared to sighted individuals across two tactile tasks. Finally, Wutte et al. (2011) report that variability in the human motion complex (as assessed by fitting a generative model) is associated with discrimination thresholds in a motion discrimination task.

Interest in BOLD variability has increased exponentially with the rise in interest in characterizing resting state activity. In most resting state studies, there is no task structure that can be used to quantify activity, which has led to the search for intrinsic markers of meaningful activity (as opposed to mere noise or background activity). One of the most widely used of these new measures is the amplitude of low-frequency fluctuations (ALFF) measure introduced by Zang et al. (2007), and its close cousin, fractional ALFF (fALFF; Zou et al. 2008), which compares the power in frequency bands putatively associated with signal against the full power spectrum. Because the spectral power of a signal is proportional to its variance, this method is in the same family as those discussed above, even if it is not always couched in the same language. It is beyond the scope of this paper to present the hundreds of results discovered with these measures, but they have each been associated with countless individual difference factors.

Concurrently with the rising interest in BOLD signal variability, the issue of motion-related artifacts has gained attention as a methodological problem. Although the bulk of the research into the influence of motion artifacts has been associated with resting state functional connectivity (e.g., Power et al. 2012, 2014; Satterthwaite et al. 2012, 2013; Van Dijk et al. 2012), motion artifacts will of course increase most measures of variability (and generally reduce data quality; Yan et al. 2013), though using a model-based approach, as in Wutte et al. (2011), may ameliorate some of this influence. Although other sources of noise—including physiological, thermal, measurement error, and so forth—are also present in fMRI, these are generally less well-studied and currently more difficult to correct for than motion.

Moreover, there may be other changes associated with, for example, aging or disease processes, that trivially influence variability (see, e.g., Kannurpatti et al 2011). Aging in particular has been associated with a number of well-documented changes to factors known to influence the BOLD signal (see, e.g., Shen et al. 2008, for a summary on the relationship of many of these factors with BOLD signal), including reduced glucose metabolism (Knopman et al. 2014); reduced cerebral metabolic rate of oxygen consumption and cerebral blood flow and increased oxygen extraction fraction (in most, but not all, of cortex; Aanerud et al. 2012); and decreased cerebral blood volume (Marchal et al. 1992).

If a researcher's goal is simply to develop a biomarker for detecting processes associated with aging or disease, then such confounds may be acceptable (in much the same way that a police department could monitor ice cream sales, rather than actual reports of crime, to determine how many officers to put on patrol). However, such confounds still increase the risk of misinterpretation—for instance, if some measure is artifactual and scanner-dependent, patients scanned elsewhere may be misdiagnosed. Moreover, if the goal is to learn something about the processes or underlying neural activity *per se*, then these confounds might completely mask the true relationships of interest.

Here, we illustrate these issues in a dataset resembling many of those used in the studies discussed above: participants aged 18–75 were scanned while they performed a recognition memory task. However, rather than presenting the empirical results, we instead demonstrate the impact of a variety of analysis choices, with a focus on attempts to control for the influence of motion. As should be clear, our intention was not an exhaustive, quantitative comparison of every possible method, but rather an illustrative demonstration of the issue using a small sampling of possible methods spanning a range of novelty and complexity. In particular, we examine three approaches for correcting for motion, representing the methodological cutting-edge (Patel et al. 2014); a variant of the most widely-used approach in the field (i.e., nuisance regression; Satterthwaite et al. 2013); and a method common in cognitive psychology more generally (i.e., partial correlation; Fisher 1924). The empirical effect we chose to focus on is the relationship between age and variability, which is among the most studied effects in the fledgling domain of BOLD variability research (Garrett et al. 2013).

As expected, subtle changes in preprocessing or analysis strategy can profoundly change the qualitative story told by the data—for example, from a map showing only regions of positive correlation between age and BOLD variability to one showing only regions of negative correlation. In addition to this illustrative result, we quantified the impact of motion and pipeline choice and found that—at the levels of motion we observed in our sample—the choice of how to correct for motion can influence the results to at least as large a degree as motion itself. Although this sort of sensitivity to choices should be familiar to researchers in the fMRI field (see, e.g., Carp 2012, for an excellent exploration of this issue from a broader stance), there is an especial need to respect these issues as fMRI—and BOLD variability in particular—comes to be increasingly used as a biomarker.

Methods

Participants

A total of 126 individuals were recruited from the UCSB and greater Santa Barbara communities. However, due to technical issues, metal screening issues, claustrophobia, and attrition, 17 participants were not included in the analyses presented here. Three age groups were assessed with the final sample consisting of 37 late adolescents (19 females, 18 males; age: 18 years), 36 young adults (17 females, 19 males; M_{age} = 28.47 years; age range: 25–33 years), and 36 older adults (19 females, 17 males; M_{age} = 67.28 years; age range: 60–75 years). Most participants were right-handed and native English speakers (one young adult reported being ambidextrous; one young adult and one older adult learned English in early childhood). All participants were free of memory complaints beyond what is common to other normal individuals in their age range and had a Mini-Mental State Examination (Folstein et al. 1975) score of 27 or higher. Informed written consent was obtained from each participant prior to any experimental procedures, all of which were approved by the University of California, Santa Barbara Human Subjects Committee.

Procedure

The data presented here come from a two-day study. Day 1 involved behavioral assessments and neuropsychological testing and will not be discussed here. All structural and functional MRI scanning occurred during Day 2. The functional MRI data were collected during a recognition memory task involving criterion shifting. Prior to scanning, participants studied 153 words, presented one at a time at the center of the screen in white font on a black background for a duration of 1.5 s, and were instructed to remember as many of the words as they could, using whatever strategy they believed would best accomplish that goal.

During the test session, the 153 old/studied words were intermixed with 153 new (i.e., unstudied) words. The test session was broken down into three separate tests, each containing 51 old words and 51 new words. A separate functional MRI scan was acquired during each test. Each test cycled through alternating blocks of high- and low-target-probability conditions with 5–7 words in each probability context (9 high- and 9 low-target-probability contexts per test). 70.6 % of the words across all high-target-probability blocks were old, while 29.4 % of the words across all low-target-probability blocks were old. During test, words were presented one at a time at the center of the screen. The font color varied to denote context: blue font for one target probability context and orange font for the other context (color/probability association counterbalanced across participants). Word and fixation stimuli (a horizontally centered crosshair of the same height as the word stimuli and same color as the current context) were presented sequentially for a duration of 1600 ms, with each word being separated by 1–4 fixation trials. Participants were instructed to indicate whether each test word was old or new by using a two-button response box. They were explicitly informed that one font color indicated that a word had a 70 % likelihood of being old, while the other font color indicated that there was a 30 % likelihood, and were encouraged to use that contextual information to help guide their old/new decision. Performance results are not the focus of this paper and will not be discussed here.

fMRI data acquisition

Scans were acquired at the UCSB Brain Imaging Center using a 3T Siemens TIM Trio MRI system with a standard 12-channel head coil. Foam cushions were placed around the head to minimize head motion. Stimuli were projected on a screen behind the participant and were viewable via a mirror mounted on the head coil. Functional runs consisted of a T2*-weighted single shot gradient echo, echo-planar sequence (interleaved, 3 mm thickness, 3 × 3 mm in-plane resolution) sensitive to BOLD contrast (TR = 1.6 s; TE = 30 ms; FA = 90°) with generalized autocalibrating partially parallel acquisitions (GRAPPA). Each volume consisted of 30 slices acquired parallel to the AC-PC plane, although the angle was slightly adjusted to optimize for frontal acquisition if necessary. A total of 316 volumes were acquired for each test run, comprising 102 stimulus trials and 214 fixation trials. A high-resolution anatomical image was collected at the beginning of the scanning session for each participant using an MPRAGE sequence (TR = 2.3 s; TE = 2.98 ms; FA = 9°; 160 slices; 1.1 mm thickness). In addition to the functional and high resolution anatomical scans, diffusion-tensor imaging and resting state scans were acquired but are not included in the present analysis.

fMRI data analysis

Although our focus here is on the impact of different analysis choices (in a descriptive, qualitative sense, rather than a prescriptive, quantitative sense), there were a number of preprocessing steps common to all of the analysis pipelines. Below, we describe these common steps, followed by each of the analysis variants in turn. In addition to sharing preprocessing, each of the analyses shares a common form: after computing voxelwise variability on a

participant-by-participant basis, these variability measures were correlated (across participants) with participant age. In general, the variants were inspired by various attempts to ameliorate concerns regarding motion, although we include an uncorrected variant as well, as not all investigations of variability have attempted to control for motion (or other noise) effects. An overview of all pipelines is given in Table 1.

Initial preprocessing

Prior to all analysis pipelines, the functional data from each functional run were preprocessed separately in FEAT, including brain extraction, spatial smoothing with a 5 mm-FWHM Gaussian kernel, motion correction, and grand-mean intensity normalization. Although this last step is default (and strongly recommended for the standard general linear model analyses for which FEAT was designed), its multiplicative nature will affect both the mean and variance of the BOLD data. Therefore, differences in BOLD grand mean will result in differences in this normalizing factor, which we refer to as the "GMIN factor", for Grand Mean Intensity Normalization factor (it is equal to $\frac{10000}{\mu_{\mathrm{BOLD}}}$). That is, although the grand mean intensity normalization carried out by FEAT ensures that all participants' data have the same mean after preprocessing, the multiplicative factor used to achieve this differs, which may influence variability; nor does *not* normalizing guarantee that this problem is solved, because we do not know whether the phenomenon driving these differences in mean across participants is also artifactually affecting variance. To preview our results, we observed differences in BOLD grand mean across individuals; moreover, these differences reliably covaried with age, a point to which we return in our Discussion.

Following computation of participant-level variability maps (see below), the maps were always transformed to standard space; a single transformation matrix was generated for each functional run per participant (based on the minimally-preprocessed data), and this matrix was used by FLIRT for all of the transformations described below across every pipeline, preventing differences in registration that might drive any of the differences observed between pipelines. We chose FLIRT rather than a (possibly more appropriate) nonlinear method such as FNIRT for simplicity (and to avoid the debate surrounding nonlinear transformation methods; Klein et al. 2009).

Minimally processed variants

The first set of pipelines simply used the preprocessed data just described with no additional denoising. We did carry out a regression that included only regressors for each of the context blocks, which we convolved with FEAT's default gamma-function HRF, along with temporal derivatives. This step was taken to remove the contribution of task-driven variability to our estimates of BOLD variability, akin to the demeaning done by Garrett et al. (2010). The regression was run using FEAT, with the additional preprocessing step of temporal filtering ($\sigma = 50.0$ s); the regressors were temporally filtered to match the data. We took the residuals of this analysis as the data on which we computed our variability maps.

To generate these variability maps, we calculated the standard deviation of the BOLD timeseries in a voxelwise manner for every functional run for each participant, and then took an average across functional runs (within participants) to generate a single variability map per participant. After transforming these variability maps into standard MNI space using FLIRT (12-degree affine transform), we generated a group-level correlation map by computing voxelwise (for every voxel with a variability measure for all participants) the Spearman correlation between participant age and variability (pipeline 1a). We converted these ρ values to z values, and thresholded the resulting map using cluster-based thresholding with a z threshold of ± 2.3 and a cluster p threshold of 0.05.

In addition to this most minimal pipeline, we examined two variants that were identical up to the point of correlating variability with age across participants, but deviated thereafter. We generated the first of these variant correlation maps (pipeline 1b) by computing the partial Spearman correlation between participant age and variability, partialing

Table 1 Pipeline descriptions (see text for additional information)

Pipeline	Wavelet denoising	Nuisance regression	Confounds partialed out
1a	No	Task only	None
1b	No	Task only	GMIN factor
1c	No	Task only	GMIN factor & mean motion
2a	No	Task & nuisance	GMIN factor
2b	No	Task & nuisance	GMIN factor & mean motion
3	Yes	Task & nuisance	GMIN factor

out GMIN factor (that is, the average GMIN factor calculated by FEAT across each participant's three functional runs). These partial correlation values were again converted to z values and thresholded as above. The second variant (pipeline 1c) was identical to the previous except that the partial Spearman correlation between age and variability was calculated by partialing out both GMIN factor and motion (that is, the average across functional runs of the "mean relative motion" estimated by FLIRT per functional run); the resulting map was again converted to z values and thresholded.

Nuisance regression variants

The next set of pipelines was similar to the above, except that the preprocessed data were processed further using a nuisance regression. In particular, the preprocessed data were regressed on an extended version of the model described earlier that included motion parameters, mean CSF signal, and context regressors. The motion parameters were the six motion parameters returned by FEAT (translation and rotation) and their temporal derivatives. The CSF regressor was derived by segmenting each participant's high resolution T1 anatomical, thresholding the resulting probabilistic CSF map at 0.9, aligning and reslicing the thresholded map to match each of the participant's functional runs (the transformation matrices were generated by registering the anatomical image with each of the mean functional images using FLIRT with a 7-degree affine transform, and then these matrices were applied to the thresholded CSF map using trilinear interpolation), thresholding again at 0.9 (to correct for the fact that some of the larger voxels in functional space may have overlapped only partially with above-threshold voxels in the higher-resolution space), and calculating the mean timeseries across all remaining suprathreshold voxels. Although there are many possible choices for the exact formulation of a nuisance regression (see, e.g. Satterthwaite et al. 2013), this form was chosen to match that recommended by Patel et al. (2014), which forms the basis of our last set of pipelines (see below).

Once again, the residuals from this regression were used to compute the variability maps, by computing the voxelwise standard deviation on the residuals for each functional run for each participant. As above, the resulting maps were averaged across functional runs (within participants), and we generated final age–variability correlation maps in two ways. For the first (pipeline 2a), we computed the partial correlation between age and variability with GMIN factor alone partialed out, while for the second (pipeline 2b), we partialed out both GMIN factor and motion (as described earlier). For each of these, the resulting maps were converted to z values and thresholded as above.

Wavelet-based variant

For the final pipeline, we used the wavelet-based method of Patel et al. (2014) to denoise the data. We chose this method for several reasons: first, as described in their paper, it compares favorably with other means of "despiking" data (e.g., using thresholds on DVARS or framewise displacement; Power et al. 2012; Van Dijk et al. 2012, Satterthwaite 2013); second, it should be expected to have a more nuanced impact on variance than wholesale removal of motion TRs (i.e., because the amplitude of those TRs is reduced to exactly the mean using a traditional despiking approach, which removes their influence on the variance completely, while the amplitude will generally be attenuated but still not equal to the mean using the wavelet-based approach); and third, it represents a relatively novel solution to the problem of motion artifacts, which fulfills our goal of covering a spectrum of possible methods here. Our major results on the influence of methodology do not depend critically on the choice of any particular denoising strategy, including this one.

To apply the wavelet-based method, the preprocessed data were entered as a single four-dimensional input directly into the WaveletDespike function from the BrainWavelet Toolbox v1.1 with all defaults left unchanged. This function returns two 4D volumes per input: one representing signal, the other noise. Using the signal component, we then carried out nuisance regression per the recommendations of Patel et al. (2014), as described above, with the additional inclusion of the convolved task regressors. As before, we computed the standard deviation across the residuals of this regression, averaged across functional runs, and finally computed the age–variability correlation map by computing the partial correlation between age and variability with GMIN factor partialed out (pipeline 3a). The resulting map was converted to z values and thresholded as above.

Quantifying relative effects of motion versus pipeline

It is of course well-established that both motion and pipeline affect the final result of any analysis (Yan et al. 2013). However, the relative magnitude of the influence of each has never been established, at least as it relates to BOLD variability analyses. We therefore examined the interaction between motion and pipeline in two ways—within participants at the level of the participant-specific variability maps (i.e., prior to computing age–variability maps), and between participants at the level of the age–variability maps; in both cases, we examined the unthresholded maps, as thresholding can exaggerate dissimilarity and is dependent on the choice of a (somewhat arbitrary) threshold. Each of these analyses is described further below.

Within-participants effect of motion

In order to get the cleanest possible estimate of the influence of motion on the variability maps themselves, we computed the pairwise similarity of the variability maps from each pair of functional runs (i.e., run 1 with run 2, run 1 with run 3, and run 2 with run 3) within participants using η^2 (see, e.g., Cohen et al. 2008, for a definition of this statistic for fMRI SPMs). Because this is within-participants, many possible sources of noise are eliminated compared to a between-participants approach. For each pair of functional runs, we compared the dissimilarity across variability maps (i.e., $1 - \eta^2$) with the difference in mean relative motion (see above) between the two scans. Note that there are three sets of variability maps (forming the bases of pipelines 1a–1c, 2a–2b, and 3, respectively). For each set of variability maps, we correlated these pairwise variability map dissimilarity values against the squared difference in mean relative motion between scans.

Between-participants effect of motion

Having established the effect of motion on variability maps, we next sought to determine the impact of motion on our results at the level of our primary focus—that is, the age–variability relationship—and to compare this impact with the impact of the analysis pipelines themselves (independent of motion). To this end, we subsetted our sample in two ways, in order to attempt to isolate the influence of either pipeline or motion, and calculated the similarity of the group-level age–variability maps across subsets (again using η^2).

In order to estimate the influence of motion on our results, we subsetted our data according to each participant's mean relative motion—that is, we performed a simple median split on mean relative motion to form low- and high-motion groups. In order to maximize the influence of motion, we computed the similarity between the age–variability maps derived by pipeline 1a. After computing the age–variability maps as described above (i.e., correlating age with variability voxelwise within the subset and converting the resulting correlation coefficients to z values) for each subset, we calculated the dissimilarity between the two resulting maps as $1 - \eta^2$ for the pair.

In order to estimate the influence of pipeline, independent (as much as possible) of motion, we subsetted our data by ordering participants according to mean relative motion, and then taking every other participant to be in one or the other group. Because we have an even number of participants, we end up with equal-sized groups that differ only very slightly on mean (across the subset) motion. Then, we computed age–variability maps using all of the pipelines described in Table 1; note that because the partial correlations used in all pipelines except 1a are meant to be computed using the full range of each confound variable, we first partialed the corresponding confounds out of the full group (that is, we separately regressed age and the voxelwise variability across subjects on the relevant set of confounds, and used the residuals of each of these regressions for subsequent steps). However, rather than comparing across subsets only within pipeline, we calculated the dissimilarity (using $1 - \eta^2$) between the subsets from every pair of pipelines—for example, comparing the "even" group (participants 2, 4, 6, … in order of mean relative motion) from pipeline 1a with the "odd" group (participants 1, 3, 5, …) from pipeline 1b and vice versa, and so forth for every pair of pipelines.

Results

Confounds with age

As we described above, there are a number of factors that influence BOLD variability and covary with age, such that changes in BOLD variability may in fact reflect the influence of these confound variables. One of these, participant motion, has been studied extensively, and is nearly universally recognized as a potential problem for certain types of analyses; Fig. 1 shows the relationship between age and mean relative motion (Spearman's $\rho = 0.51$) in our sample. A second factor is a change in mean BOLD (of course, holding scanner, sequence, etc. constant) with age; Fig. 2 shows the relationship between age and GMIN factor (which, as we described above, is inversely proportional to the BOLD

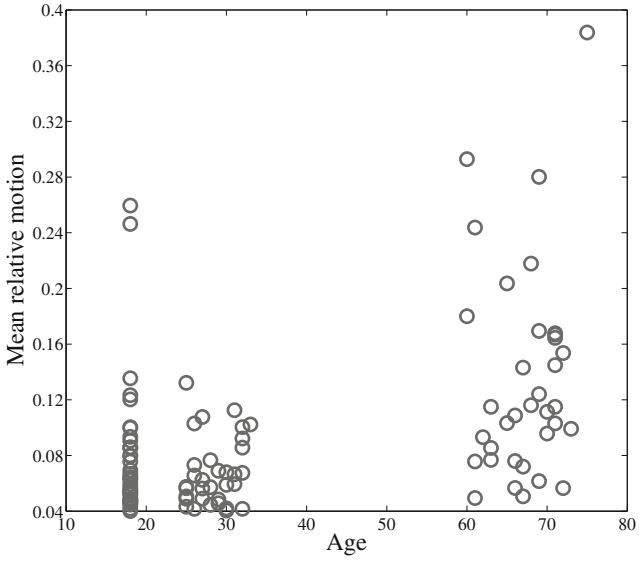

Fig. 1 Scatterplot demonstrating the relationship between age and mean relative motion. The distinct banding on the age axis is due to the discontinuity in age ranges for our three groups

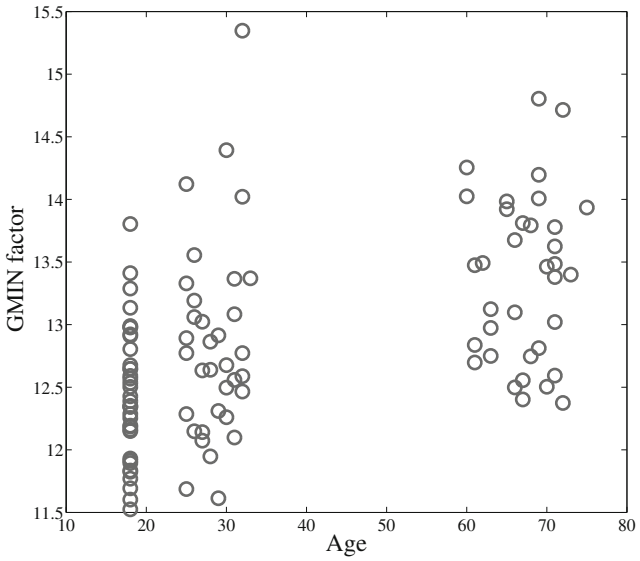

Fig. 2 Scatterplot demonstrating the relationship between age and GMIN factor

grand mean—that is, GMIN factor increases with age, while BOLD grand mean decreases with age; Spearman's $\rho = 0.43$).

Analysis variant results

Below, we present the rendered thresholded maps yielded by each of the analysis variants described above. As a reminder, the threshold was held constant, and there were no changes made to the baseline—that is, all differences are due to the impact of the choices made in each pipeline.

Figure 3 shows the results obtained with the minimal pipeline (1a) without partialing out any possible confounds. Note that there are extensive regions of positive correlation between age and variability. When GMIN factor is partialed out (pipeline 1b), the resulting map is as shown in Fig. 4; now, many of the regions that were suprathreshold in the uncorrected correlation are subthreshold, yielding a substantially sparser map. However, the most drastic difference for this pipeline comes when motion is additionally partialed out (pipeline 1c), resulting in the map shown in Fig. 5.

When nuisance regression is applied prior to computing the correlation maps, the results are broadly similar to the minimal results, particularly when motion is partialed out. Figure 6 shows the analogous map to Fig. 4, with nuisance regression and partialing out GMIN factor (pipeline 2a). The positive regions are very similar between these two, but the nuisance regression has revealed a set of regions of negative correlation. Figure 7 likewise shows the nuisance regression analogue to Fig. 5, partialing out both GMIN factor and mean relative motion (pipeline 2b). Both figures show extensive areas of negative correlation, and the differences between the figures are minimal.

Finally, when wavelet despiking and nuisance regression are both applied prior to computing the correlation maps, the results are as shown in Fig. 8. A pattern similar to some of those observed previously obtains here: Fig. 8 reveals regions of positive and negative correlation, in line with Fig. 6.

Effects of motion on within-participant variability maps

There was a strong, significant relationship between the squared difference in mean relative motion between functional runs and the dissimilarity between the resulting

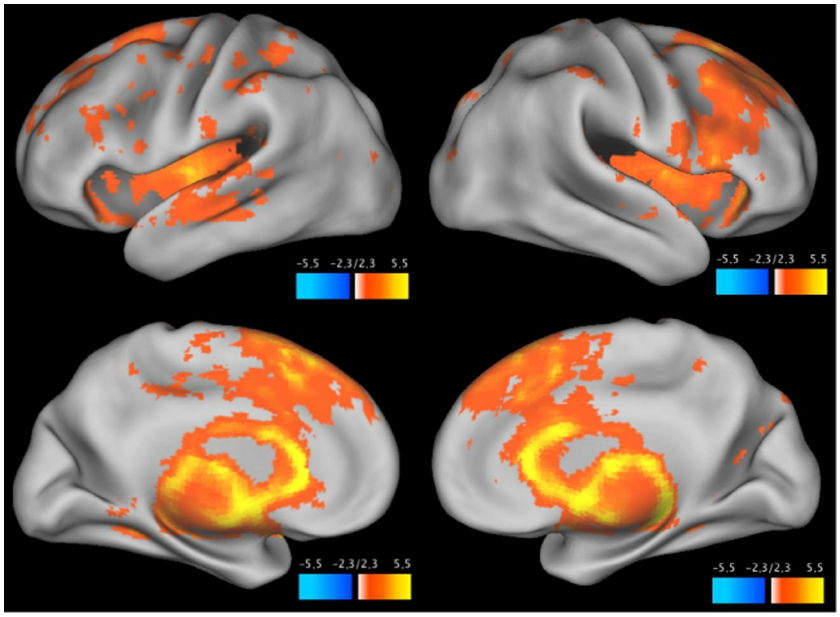

Fig. 3 Thresholded correlation map between age and variability using the minimal processing pipeline and partialing out no confounds (pipeline 1a)

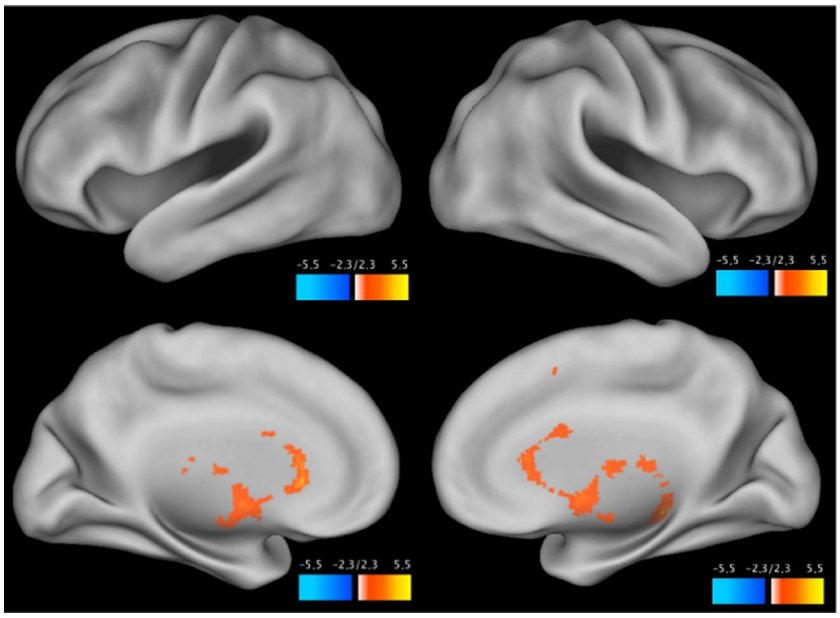

Fig. 4 Thresholded correlation map between age and variability using the minimal processing pipeline with GMIN factor partialed out (pipeline 1b)

variability maps on a within-participants basis. For the minimally-processed variability maps (those that were used in pipelines 1a–1c), the Pearson correlation between the squared difference in mean relative motion for a pair of scans and the dissimilarity of the corresponding variability maps was $r = 0.90$, $p < 0.001$ (all p-values calculated using dfs equal to the number of participants to conservatively correct for nonindependence introduced by each participant having contributed three values); removal of the highest-leverage participant still resulted in a correlation of $r = 0.72$, $p < 0.001$.

The correlations for the other two sets of variability maps are similar; for the nuisance-regression variability maps (used for pipelines 2a–2b), the correlation was $r = 0.92$, $p < 0.001$ ($r = 0.78$, $p < 0.001$ with highest-leverage participant removed), and for the wavelet-based variability maps (used for pipeline 3), the correlation was $r = 0.92$, $p < 0.001$ ($r = 0.75$, $p < 0.001$ with highest-leverage participant removed).

However, using the correlation coefficient masks some of the influence of pipeline on the form of the motion–dissimilarity relationship, in particular the slope and intercept. Therefore, we repeated each of the above using a simple regression analysis to quantify whether the different analysis methods in fact changed the influence of motion. Because we are not adding any regressors, we will not

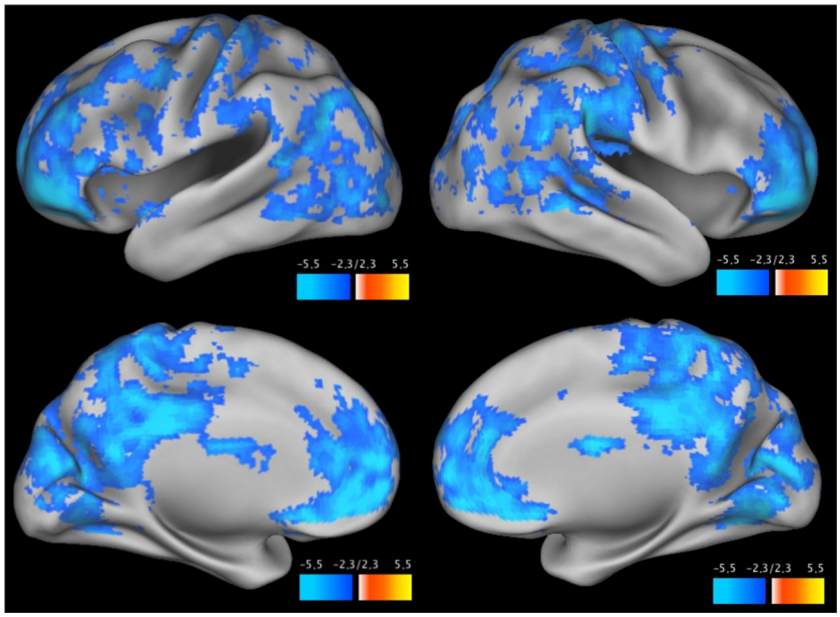

Fig. 5 Thresholded correlation map between age and variability using the minimal processing pipeline with both GMIN factor and mean relative motion partialed out (pipeline 1c)

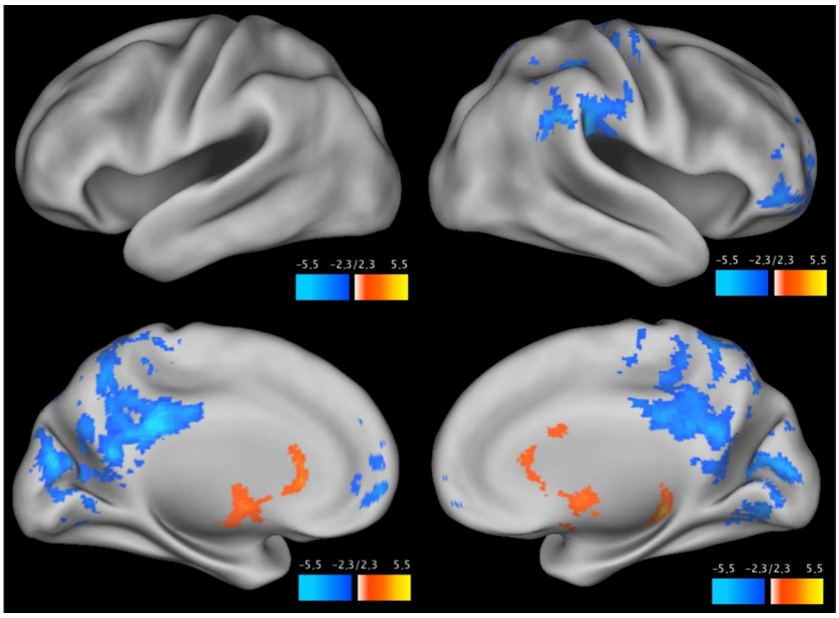

Fig. 6 Thresholded correlation map between age and variability using the nuisance regression (but no wavelet despiking) pipeline with GMIN factor partialed out (pipeline 2a)

restate the inferential statistics for each of the slope betas (which will be identical to the corresponding correlation *p*-values given above). For the minimal pipeline, the slope between squared mean relative motion difference (scaled by 100) and dissimilarity was 0.048—that is, every 0.01 unit change in squared mean relative motion difference results in an increase in dissimilarity of 0.048—with an intercept of 0.010. For the nuisance-regression pipeline, this slope was 0.038, with an intercept of 0.006; and for the wavelet-based pipeline, this slope was 0.033, with an intercept of 0.009.

Across the sets of maps, we therefore see that the strength of the relationship is approximately equal (all *r*-values between 0.90 and 0.92, or 0.72 and 0.78 using the more conservative estimate). However, the slope of this relationship—that is, the amount of additional dissimilarity per unit of additional squared difference in mean motion—gets progressively shallower across methods, while the intercept—that is, the dissimilarity between maps with identical amounts of motion—stays almost constant near 0.01.

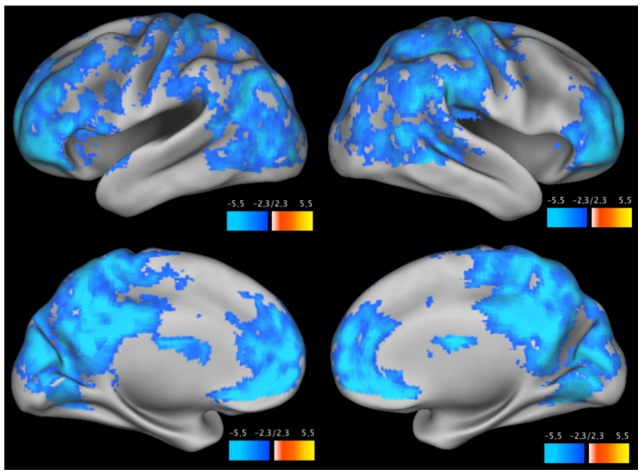

Fig. 7 Thresholded correlation map between age and variability using the nuisance regression pipeline with both GMIN factor and mean relative motion partialed out (pipeline 2b)

Effects of motion and pipeline on age–variability maps

When we split our sample in low- and high-motion subsets, our low-motion subgroup has a mean relative motion (across the subgroup) of 0.054 (standard deviation = 0.01), and our high-motion subgroup has a mean of 0.134 (sd = 0.07), for a difference in mean motion between groups of 0.080. Using the approach described in the Methods—that is, separately computing age–variability maps in each subgroup, then computing the dissimilarity between the two—we get a dissimilarity of 0.2731. Although we cannot quantify the impact of differing amounts of motion as we did for the within-participant analysis, we will use 0.2731 as a benchmark against which to compare the impact of pipeline.

When we instead split our sample to approximately match motion, our "odd" subgroup has a mean relative motion of 0.092 (sd = 0.06), while our "even" subgroup has a mean of 0.096 (sd = 0.07), indicating that we achieved our goal (i.e., the group means differ by only 0.004 in motion). The results of each pairwise dissimilarity computation (averaged across both orderings of the pipeline pair, i.e., "odd" pipeline A with "even" pipeline B and vice versa, under the presumption of symmetry) are presented in Table 2; because the dissimilarity between subgroups *within* any given pipeline varies substantially, we also present "corrected" dissimilarities (above the diagonal), which we

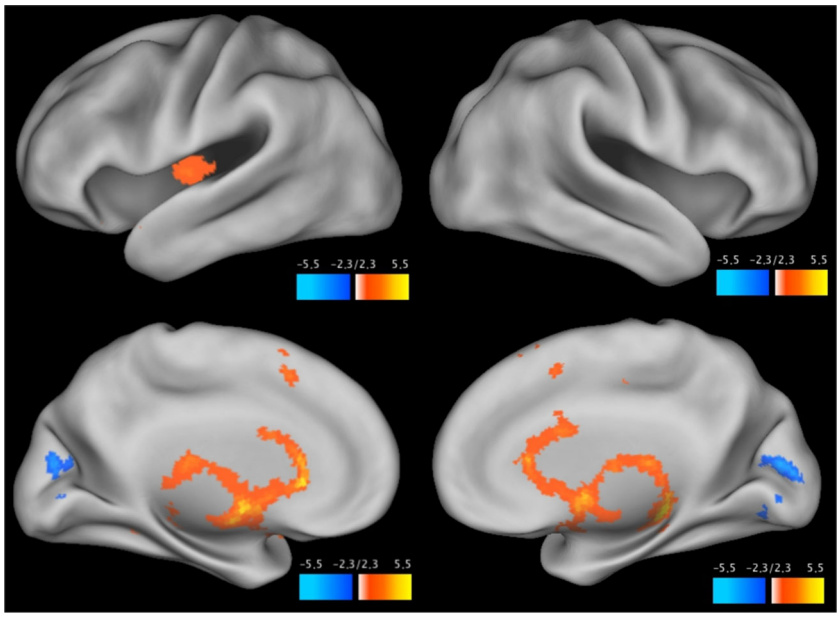

Fig. 8 Thresholded correlation map between age and variability using the wavelet despiking pipeline with GMIN factor partialed out (pipeline 3)

generated by subtracting the average within-pipeline dissimilarity for each pipeline-pair from the between-pipeline dissimilarity for the pair. This removes the intrinsic dissimilarity associated with any particular pipeline and leaves only the dissimilarity added by comparing *across* pipelines.

Note that compared to the dissimilarity observed in our low- vs. high-motion analysis, the between-pipeline dissimilarities range from almost zero to much higher than the dissimilarity due to motion. Moreover, the within-pipeline dissimilarity for pipeline 1a (which we used to estimate dissimilarity in the low- vs. high-motion analysis) is only 0.0271 in our analysis of the effect of pipeline, consistent with our intuition that a lower motion difference between groups (0.004 rather than 0.080) should result in lower dissimilarity.

Discussion

Our results demonstrate several key ideas. Perhaps most importantly, they reveal that there is some consistent structure to the patterns in our data. Across the (admittedly finite) range of variants we tested, there were ensembles of regions which, when suprathreshold, always evinced either a positive or negative correlation between variability and age—for instance, when it appears on a map, precuneus always shows a negative correlation, and likewise, the subcortical regions visible on several maps always have a positive correlation. Therefore, our point is not that there is no underlying truth, or that our results are so flexible that the endeavor is fruitless. However, it is also clear that there is some ambiguity as to the *exact* truth.

This second idea, of course, is our broader point. Any researcher who used one of these pipelines, in ignorance of the other possible results, might simply accept whichever map came out at the end. Unfortunately, as demonstrated above, the qualitative picture of how variability relates to age depends critically on the exact pipeline used. Some of the pipelines are more defensible than others, but even two that on the face seem to have very similar goals and to be reasonable—e.g., Figs. 7 and 8, which take two different approaches to dealing with residual motion artifacts not corrected by nuisance regression—yield substantially different conclusions. In one case, there is extensive negative

Table 2 Dissimilarity in age–variability maps between pipelines. Values above diagonal (in **bold**) denote corrected values, while values on diagonal and below denote uncorrected values; for example, the corrected dissimilarity between pipelines 1a and 1b is .0285, which is equal to $.1410 - \frac{.0271+.1980}{2}$ (i.e., the uncorrected between-pipeline dissimilarity minus the average of the within-pipeline dissimilarities)

	Pipeline 1a	Pipeline 1b	Pipeline 1c	Pipeline 2a	Pipeline 2b	Pipeline 3
Pipeline 1a	.0271	**.0285**	**.6771**	**.6863**	**.6735**	**.0270**
Pipeline 1b	.1410	.1980	**.1634**	**.1603**	**.1598**	**.0080**
Pipeline 1c	.7017	.2736	.0221	**-.0001**	**.0015**	**.2068**
Pipeline 2a	.7143	.2737	.0254	.0288	**.0027**	**.2090**
Pipeline 2b	.6952	.2670	.0207	.0252	.0163	**.2004**
Pipeline 3	.1264	.1929	.3037	.3092	.2944	.1717

correlation between the two variables, with no regions showing the opposite pattern, while in the other, there is approximately equal representation of brain areas showing positive and negative correlations, with much less of the brain showing any pattern at all.

There are several aspects of the results from our quantification of the relative impacts of motion and pipeline that are worth highlighting. From the within-participant analyses, perhaps the most striking result is the high degree of consistency across functional runs in the presence of low amounts of motion. At the limit of no difference in motion between the scans, the amount of dissimilarity between variability maps for all three methods we tested was 0.01 or below (and roughly equal across methods, suggesting that none of the methods overcorrects or introduces dissimilarity in the absence of motion). Granted, this is a very high-level measure, and any particular area of the brain may in fact be much less consistent across functional runs, but the whole-brain pattern within an individual (at the very short delay interval we used in this study) is markedly stable. Moreover, motion exerts an extremely reliably influence on dissimilarity, accounting for approximately 52–61 % of the variance in dissimilarity measures even after removing outliers, corresponding to an increase of 0.03–0.05 in dissimilarity for every increase of 0.01 in squared difference in mean relative motion. Lastly, the pipelines seem to be at least partially achieving their goals of ameliorating motion's influence, as each successive processing step reduced the slope of that relationship.

Turning to our analysis of the impact of motion and pipeline at the level of age–variability maps, we found that the dissimilarity between the age–variability maps of the low- and high-motion subsets of our sample was 0.27; nor was the difference in motion especially larger or smaller than would be expected in any study that included a cohort known to have high motion (in our case, older adults): the two groups differed in mean relative motion by 0.08. On the one hand, it is encouraging to note (as we pointed out qualitatively above) that the age–variability maps are still fairly similar, even when trying to maximize dissimilarity due to motion. However, the more worrying result is the dissimilarity between pipelines, when trying to minimize the influence of motion. There, across pairs of pipelines, we see corrected dissimilarities ranging between 0.00 and 0.69, with a median corrected dissimilarity of 0.16, and a range of 0.69 between the most and least dissimilar pairs of pipelines. In other word, at the levels of motion observed in our sample, the choice of pipeline exerts a (potentially substantially) larger impact on the resulting age–variability maps than motion itself.

One other result that warrants additional attention is the relationship between GMIN factor and age. It is often the case that for real-world Gaussian-distributed (or approximately Gaussian-distributed) variables, changes in mean are accompanied by changes in variance. When the BOLD grand mean differs between participants systematically, any differences in variance may therefore be artifactual, such that the grand-mean intensity normalization corrects for this artifact. However, it may be that such differences in variance (even if they accompany differences in mean) are "true" differences, in which case the multiplicative scaling in fact masks the truth. Here, we chose to perform the normalization, but to partial out the GMIN factor in most of our analyses. We stress that this phenomenon is distinct from FEAT or use of grand-mean intensity normalization—rather, we observed that mean BOLD varies with age, and we are simply using the GMIN factor as a proxy for this phenomenon, to correct for its influence on variability. However, this effect requires additional investigation, as it was not the focus of the present study.

Considering that we included only a small sample of closely related analysis streams, it should be evident that there is considerable uncertainty regarding what the true pattern is. The issue is even more complicated when denoising methods that allow some degree of subjectivity—e.g., ICA-based denoising with manual selection of noise components—are included, or if researchers try multiple analyses and choose the "right" one *post hoc*, unblind to the results. For variability analyses in particular, there is the additional complicating issue of global mean BOLD varying reliably with age; although we are agnostic to whether accompanying changes in variability are ephiphenomenal, this is an effect which, to our knowledge, has not been previously described, and which should be considered in analyses of BOLD variability.

It is unlikely that additional analysis steps (e.g., partial least squares) or different choices of individual covariates (e.g., behavioral measures rather than age) would ameliorate the concerns raised here. That is, the inconsistency we observe across pipelines does not reflect the exact choices we made here, in terms of the relationships we investigated, or the exact measures used to investigate them. Indeed, our point is that this flexibility is an unavoidable property of these sorts of analyses, and that researchers should carefully consider any choices they make. Moreover, although we only demonstrated a relationship between mean BOLD and age here, there may be other variables with which mean BOLD covaries.

Ultimately, these issues become more important as research takes on a more applied focus. Although there are theoretical ramifications to drawing incorrect conclusions—for instance, whether variability increases with age and reflects compensatory mechanisms (as suggested in a different modality by, e.g., Cabeza et al. 1997, 2002), or rather decreases with age and reflects loss of flexibility (as suggested by, e.g., Garrett et al. 2011)—the stakes are

higher when these theoretical constructs are appropriated as biomarkers. If high (or low) variability in some region is used as an indicator of pathology—and particularly if the exact method by which that variability is computed is not applied canonically—the possibility of subjecting individuals to needless worry or unnecessary treatments becomes almost a certainty. Therefore, it is critical that any result considered for applied use be highly robust and repeatable—which, given the disagreement in the literature and the inconsistency demonstrated here, BOLD variability's relationship with age clearly is not.

Acknowledgments This work was supported by the Institute for Collaborative Biotechnologies through contract no. W911NF-09-D-0001 from the U.S. Army Research Office.

Conflict of Interest Benjamin O. Turner, Brian Lopez, Tyler Santander, and Michael B. Miller declare that they have no conflicts of interest.

Informed Consent All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, and the applicable revisions at the time of the investigation. Informed consent was obtained from all patients for being included in the study.

References

Aanerud, J., Borghammer, P., Chakravarty, M.M., Vang, K., Rodell, A.B., Jónsdottir, K.Y., Møller, A., Ashkanian, M., Vafaee, M.S., Iversen, P., et al. (2012). Brain energy metabolism and blood flow differences in healthy aging. *Journal of Cerebral Blood Flow & Metabolism*, *32*(7), 1177–1187.

Cabeza, R., Grady, C.L., Nyberg, L., McIntosh, A.R., Tulving, E., Kapur, S., Jennings, J.M., Houle, S., Craik, F.I. (1997). Age-related differences in neural activity during memory encoding and retrieval: a positron emission tomography study. *The Journal of Neuroscience*, *17*(1), 391–400.

Cabeza, R., Anderson, N.D., Locantore, J.K., McIntosh, A.R. (2002). Aging gracefully: compensatory brain activity in high-performing older adults. *Neuroimage*, *17*(3), 1394–1402.

Carp, J. (2012). The secret lives of experiments: methods reporting in the fmri literature. *Neuroimage*, *63*(1), 289–300.

Cohen, A.L., Fair, D.A., Dosenbach, N.U., Miezin, F.M., Dierker, D., Van Essen, D.C., Schlaggar, B.L., Petersen, S.E. (2008). Defining functional areas in individual human brains using resting functional connectivity mri. *Neuroimage*, *41*(1), 45–57.

Fisher, R.A. (1924). The distribution of the partial correlation coefficient. *Metron*, *3*, 329–332.

Folstein, M.F., Folstein, S.E., McHugh, P.R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Garrett, D.D., Kovacevic, N., McIntosh, A.R., Grady, C.L. (2010). Blood oxygen level-dependent signal variability is more than just noise. *The Journal of Neuroscience*, *30*(14), 4914–4921.

Garrett, D.D., Kovacevic, N., McIntosh, A.R., Grady, C.L. (2011). The importance of being variable. *The Journal of Neuroscience*, *31*(12), 4496–4503.

Garrett, D.D., Kovacevic, N., McIntosh, A.R., Grady, C.L. (2012). The modulation of bold variability between cognitive states varies by age and processing speed. *Cerebral Cortex*, 684–693.

Garrett, D.D., Samanez-Larkin, G.R., MacDonald, S.W., Lindenberger, U., McIntosh, A.R., Grady, C.L. (2013). Moment-to-moment brain signal variability: a next frontier in human brain mapping *Neuroscience & Biobehavioral Reviews*, *37*(4), 610–624.

He, B.J. (2011). Scale-free properties of the functional magnetic resonance imaging signal during rest and task. *The Journal of neuroscience*, *31*(39), 13,786–13,795.

Kannurpatti, S.S., Motes, M.A., Rypma, B., Biswal, B.B. (2011). Increasing measurement accuracy of age-related bold signal change: Minimizing vascular contributions by resting-state-fluctuation-of-amplitude scaling. *Human Brain Mapping*, *32*(7), 1125–1140.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jerkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Ramin, V.P. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*, *46*(3), 786–802.

Knopman, D.S., Jack Jr, C.R., Wiste, H.J., Lundt, E.S., Weigand, S.D., Vemuri, P., Lowe, V.J., Kantarci, K., Gunter, J.L., Senjem, M.L., et al (2014). ^{18}F-fluorodeoxyglucose positron emission tomography, aging, and apolipoprotein e genotype in cognitively normal persons. *Neurobiology of Aging*, *35*, 2096–2106.

Leo, A., Bernardi, G., Handjaras, G., Bonino, D., Ricciardi, E., Pietrini, P. (2012). Increased bold variability in the parietal cortex and enhanced parieto-occipital connectivity during tactile perception in congenitally blind individuals. *Neural plasticity*, 2012.

Liu, C.Y., Krishnan, A.P., Yan, L., Smith, R.X., Kilroy, E., Alger, J.R., Ringman, J.M., Wang, D.J. (2013). Complexity and synchronicity of resting state blood oxygenation level-dependent (bold) functional mri in normal aging and cognitive decline. *Journal of Magnetic Resonance Imaging*, *38*(1), 36–45.

Marchal, G., Rioux, P., Petit-Taboué, M.C., Sette, G., Travere, J.M., Le Poec, C., Courtheoux, P., Derlon, J.M., Baron, J.C. (1992). Regional cerebral oxygen consumption, blood flow, and blood volume in healthy human aging. *Archives of neurology*, *49*(10), 1013–1020.

Mohr, P.N., & Nagel, I.E. (2010). Variability in brain activity as an individual difference measure in neuroscience *The Journal of Neuroscience*, *30*(23), 7755–7757.

Patel, A.X., Kundu, P., Rubinov, M., Jones, P.S., Vértes, P.E., Ersche, K.D., Suckling, J., Bullmore, E.T. (2014). A wavelet method for modeling and despiking motion artifacts from resting-state fmri time series. *NeuroImage*, *95*, 287–304.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E. (2012). Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *Neuroimage*, *59*(3), 2142–2154.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fmri. *Neuroimage*, *84*, 320–341.

Samanez-Larkin, G.R., Kuhnen, C.M., Yoo, D.J., Knutson, B. (2010). Variability in nucleus accumbens activity mediates age-related suboptimal financial risk taking. *The Journal of Neuroscience*, *30*(4), 1426–1434.

Satterthwaite, T.D., Wolf, D.H., Loughead, J., Ruparel, K., Elliott, M.A., Hakonarson, H., Gur, R.C., Gur, R.E. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage*, *60*(1), 623–632.

Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H.,

Gur, R.C., Gur, R.E., et al (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, *64*, 240–256.

Shen, Q., Ren, H., Duong, T.Q. (2008). Cbf, bold, cbv, and cmro2 fmri signal temporal dynamics at 500-msec resolution. *Journal of Magnetic Resonance Imaging*, *27*(3), 599–606.

Van Dijk, K.R., Sabuncu, M.R., Buckner, R.L. (2012). The influence of head motion on intrinsic functional connectivity mri. *Neuroimage*, *59*(1), 431–438.

Wutte, M.G., Smith, M.T., Flanagin, V.L., Wolbers, T. (2011). Physiological signal variability in hmt+ reflects performance on a direction discrimination task. *Frontiers in Psychology*, *2*.

Yan, C.G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R.C., Martino, A.D., Li, Q., Zuo, X.N., Castellanos, F.X., Milham, M.P. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *NeuroImage*, *76*(0), 183–201.

Zang, Y.F., He, Y., Zhu, C.Z., Cao, Q.J., Sui, M.Q., Liang, M., Tian, L.X., Jiang, T.Z., Wang, Y.F. (2007). Altered baseline brain activity in children with adhd revealed by resting-state functional mri. *Brain and Development*, *29*(2), 83–91.

Zou, Q.H., Zhu, C.Z., Yang, Y., Zuo, X.N., Long, X.Y., Cao, Q.J., Wang, Y.F., Zang, Y.F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (alff) for resting-state fmri: fractional alff. *Journal of Neuroscience Methods*, *172*(1), 137–141.