Optimizing Preprocessing and Analysis Pipelines for Single-Subject FMRI. I. Standard Temporal Motion and Physiological Noise Correction Methods

Nathan W. Churchill,[1,2]* Anita Oder,[1] Hervé Abdi,[3] Fred Tam,[1]
Wayne Lee,[4] Christopher Thomas,[5] Jon E. Ween,[6,7]
Simon J. Graham,[1,2,8] and Stephen C. Strother[1,2]

[1]*Rotman Research Institute, Baycrest, Toronto, Ontario, Canada*
[2]*Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada*
[3]*School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas*
[4]*Diagnostic Imaging, Hospital for Sick Children, Toronto, Ontario, Canada*
[5]*Nova Scotia Cancer Center, Halifax, Nova Scotia, Canada*
[6]*Posluns Centre for Stroke and Cognition, Kunin-Lunenfeld Applied Research Unit, Baycrest, Toronto, Ontario, Canada*
[7]*Division of Neurology, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada*
[8]*Imaging Research, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada*

Abstract: Subject-specific artifacts caused by head motion and physiological noise are major confounds in BOLD fMRI analyses. However, there is little consensus on the optimal choice of data preprocessing steps to minimize these effects. To evaluate the effects of various preprocessing strategies, we present a framework which comprises a combination of (1) nonparametric testing including reproducibility and prediction metrics of the data-driven NPAIRS framework (Strother et al. [2002]: NeuroImage 15:747–771), and (2) intersubject comparison of SPM effects, using DISTATIS (a three-way version of metric multidimensional scaling (Abdi et al. [2009]: NeuroImage 45:89–95). It is shown that the quality of brain activation maps may be significantly limited by sub-optimal choices of data preprocessing steps (or "pipeline") in a clinical task-design, an fMRI adaptation of the widely used Trail-Making Test. The relative importance of motion correction, physiological noise correction, motion parameter regression, and temporal detrending were examined for fMRI data acquired in young, healthy adults. Analysis performance and the quality of activation maps were evaluated based on Penalized Discriminant Analysis (PDA). The relative importance of different preprocessing steps was assessed by (1) a nonparametric Friedman rank test for fixed sets of preprocessing steps, applied to all subjects; and (2) evaluating pipelines chosen specifically for each subject. Results demonstrate that preprocessing choices have significant, but subject-dependant effects, and that individually-optimized pipelines may significantly improve the reproducibility of fMRI results over fixed pipelines. This was demonstrated by the detection of a significant interaction with motion parameter regression and physiological noise

Contract grant sponsor: CIHR; Contract grant number: MOP84483; Contract grant sponsor: HSF; Contract grant number: NA6391; Contract grant sponsors: James S. McDonnell Foundation, NSERC, Baycrest Neurovascular Fellowship, Heart and Stroke Foundation of Ontario (Centre for Stroke Recovery).

*Correspondence to: Nathan W. Churchill, Rotman Research Institute, Baycrest, 3560 Bathurst Street, Toronto, ON.
E-mail: nchurchill@rotman-baycrest.on.ca

Received for publication 18 June 2010; Revised 6 November 2010; Accepted 18 November 2010

DOI: 10.1002/hbm.21238

Published online 31 March 2011 in Wiley Online Library (wileyonlinelibrary.com).

correction, even though the range of subject head motion was small across the group ($\ll$ 1 voxel). Optimizing pipelines on an individual-subject basis also revealed brain activation patterns either weak or absent under fixed pipelines, which has implications for the overall interpretation of fMRI data, and the relative importance of preprocessing methods. *Hum Brain Mapp 33:609–627, 2012.* © 2011 Wiley Periodicals, Inc.

Key words: BOLD fMRI; preprocessing; model optimization; data-driven metrics; head motion; physiological noise; multivariate analysis

INTRODUCTION

Functional magnetic resonance imaging (fMRI), an invaluable tool for the non-invasive analysis of brain function, is used in experimental as well as clinical neuroscience contexts. However, fMRI data have low contrast-to-noise ratio (CNR), and are characterized by complex spatiotemporal signal and noise patterns. To alleviate these two problems, a vast array of preprocessing methods and analysis techniques are applied to raw fMRI data prior to data analysis. Most fMRI analyses are performed under the implicit assumptions that either the results are relatively insensitive to the chosen set of preprocessing steps (e.g. the preprocessing pipeline), or that the default settings in established software packages give near-optimal results. In recent years, though, it has been repeatedly shown that both standard and non-standard data preprocessing choices may have significant effects (negative or positive) on the quality of extracted results (e.g., [Della-Maggiore et al., 2002; Kay et al., 2007; Morgan et al., 2007; Murphy et al., 2009; Poline et al., 2006; Sarty, 2007; Strother et al., 2004; Tanabe et al., 2002; Zhang et al., 2009]. To draw reliable conclusions from fMRI results, it is thus necessary to evaluate the interactions of preprocessing and data analysis choices in a rigorous, systematic manner.

The existing literature is divided over the correct temporal preprocessing choices used to correct for subject head movement and physiological noise. Many studies demonstrate the importance of standard motion correction (MC) algorithms [Ardekani et al., 2001; Friston et al., 1995a,b; Jiang et al., 1995; Morgan et al., 2007; Oakes et al., 2005]. Nonetheless, standard rigid-body alignment may introduce significant artifacts, which may be corrected by non-rigid alignment methods [Bannister et al., 2004; Kim et al., 1999]. Freire and Mangin [2001] and Orchard and Atkins [2003] also demonstrated that the commonly-used least-squares cost functions may be susceptible to activation biases. Another common motion correction technique, termed motion parameter regression (MPR), regresses out signal temporally correlated with head movement. The effectiveness of this technique is also disputed: MPR has been found to both improve reliability of fMRI results [Freire and Mangin, 2001], and reduce noise variance [Lund et al., 2005], particularly in young children compared to adults [Evans et al., in press]. However, other research has shown that applying MPR in block designs reduces the strength of task activation [Johnstone et al., 2006], and that the procedure may remove both fMRI signal and artifact indiscriminately [Ollinger et al., 2009]. It has also been suggested that such techniques are influenced by patterns of head motion; for example, it has been shown that MPR is a more important denoising step in cases of task-correlated motion [Bullmore et al., 1999; Johnstone et al., 2006], while Freire and Mangin [2001] anecdotally report that MC may be more influenced by task-activation bias in cases of low head motion.

Physiological noise correction (PNC) has been generally shown to improve results [Glover et al., 2001; Hu et al., 1995], however, additional regression may also be required to remove physiological artifacts effectively [Birn et al., 2006; Chang et al., 2009]. Additionally, Jones et al. [2008] have shown that the effectiveness of PNC depends on its order in the preprocessing pipeline. Low-order temporal detrending has also been found to significantly affect results, depending on choice of detrending basis [Kay et al., 2007; Tanabe et al., 2002] and interactions with other parameters, such as spatial smoothing [Shaw et al., 2003]. A number of studies have also investigated basis decomposition techniques using Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for data denoising [e.g. McKeown et al., 1998; Thomas et al., 2002]. These methods have been generally found to improve signal detection, but there is little consensus on how effectively they separate signal from noise.

To test the effects of preprocessing choice on fMRI results, pipeline performance should ideally be directly measured from experimental data. Because there is no "ground truth" in fMRI data (e.g., which activations are task-related and which are artifact), techniques such as the data-driven framework of NPAIRS (Nonparametric Prediction, Activation, Influence, and Reproducibility reSampling) are particularly important for evaluating pipeline performance [Strother et al., 2002, 2010]. NPAIRS uses cross-validation data resampling to validate results, generating metrics of spatial reproducibility and temporal prediction accuracy. Reproducibility quantifies the global similarity of statistical parametric maps (SPMs), and is correlated with global signal-to-noise (SNR) of the SPMs. Prediction measures how consistently the estimated model parameters from a training dataset can predict the experimental condition for each brain scan of another independent test dataset.

The present work extends the results of LaConte et al. [2003], Shaw et al. [2003], Strother et al. [2004], and Zhang et al. [2008, 2009], among others, who showed that spatial smoothing, temporal filtering and motion correction have significant impacts on fMRI results, as measured with NPAIRS metrics. These results also revealed measurable subject heterogeneity, in which different sets of preprocessing choices performed better for different subjects. Therefore, these results are herein extended to assess commonly used temporal-based preprocessing methods for correcting subject artifact, and to examine both their effects in isolation and interactions, in fMRI of a clinically-relevant behavioral task. Specifically, the NPAIRS metrics were used to test motion and physiological noise correction along with temporal detrending, performed with standard neuroimaging preprocessing tools. To evaluate these preprocessing effects, this article presents a combination of nonparametric testing using NPAIRS metrics, and DISTATIS (a three-way multidimensional scaling technique [Abdi et al., 2009]), which may be used as a pipeline testing framework.

Analyses were performed using Penalized Discriminant Analysis (PDA) built on an optimized subset of principal components (PCs) [Yourganov et al., in press]. The PDA model is a multivariate testing adaptation of linear discriminant analysis which, unlike univariate methods, such as the General Linear Model (GLM), incorporates full brain volume spatio-temporal variance into the analysis model. PDA is able to extract distributed activation networks in the brain [Moeller and Strother, 1991; Friston et al., 1995b], but may be more sensitive to subject-dependent artifacts. For example, Zhang et al. [2008, 2009] showed that motion correction in a motor task had a large effect on PDA results, whereas GLM results were less affected. The PDA analyses were performed on fMRI data acquired during administration of the Trail-Making Test, widely used in clinical neuropsychology [Hachinski et al., 2006; Stuss et al., 2001]. This task forms part of an fMRI clinical assessment battery that is under development for assessing brain activation in stroke patients, and consists of short acquisitions of less than 3 min per run. An important outcome is that, with optimal preprocessing choices, such rapid task sets, using asynchronous motor responses, can provide strongly predictive, reliable results for individual subjects.

For the tested experimental design, it is demonstrated that (1) preprocessing choices have significant, common effects on performance, across all subjects; however, (2) there is a heterogeneous set of subject-specific pipelines that significantly improve on the optimal fixed-pipeline choice. In addition, (3) a driving effect of between-subject pipeline heterogeneity is the range of subject head movement; even for a relatively modest motion range (e.g., $\ll 1$ voxel), there are significant, systematic differences in pipeline effect. Finally, (4) individual-subject optimization tends to increase the spatial extent of shared task activations, but with a trade-off of increased between-subject variance in voxel-wise signal amplitude of the SPM activation regions. From the results of (4), it is shown that the chosen method of optimization and associated metrics may significantly affect the pattern and spatial extent of extracted activation maps.

METHODS

Data Acquisition

The experimental task was an adaptation of the Trail-Making Test [Army Individual Test Battery, 1944; Stuss et al., 2001], designed for the fMRI environment [Tam et al., 2010]. Task blocks alternately consisted of Trails A, where numbers 1–14 were pseudorandomly distributed on a viewing screen, and Trails B, where numbers 1–7 and letters A–G were shown. Subjects drew a line connecting items in sequence (1-2-3-... or 1-A-2-B-...) as quickly as possible while maintaining accuracy, over a 20 s block. After each task block, a 20-s baseline block was shown, in which subjects drew a line from the center of the screen to a dot (randomly placed at fixed radius from the center) and back, once every 2 s. A 4-block, 40-scan epoch of Trails A-Baseline-Trails B-Baseline was performed two times per run (80 scans for 160 s/run), with two runs per subject. Subjects performed line drawing tasks with an fMRI-compatible writing tablet and stylus, and monitored their performance on a projection screen [Tam et al., 2010]. Fifteen young, healthy volunteers (eight females) participated in the study, Ages 20–32 years with mean age 25.2 years. Subjects were confirmed to be right-handed using the Edinburgh Handedness Inventory [Oldfield et al., 1971], and screened for cognitive and neurological deficits, by self-report and using the Mini-Mental Status Examination [Folstein et al., 1975], with a group mean of 29.7 ± 0.6 (out of 30).

The BOLD fMRI data were acquired on a 3T MR scanner (MAGNETOM Tim Trio, VB15A software; Siemens AG, Erlangen, Germany), with a 12-channel head coil. A T1-contrast anatomical scan was obtained (oblique-axial 3D MPRAGE, 2.63/2,000/1,100 ms TE/TR/TI, 9° FA, 256 × 192 matrix, 160 slices per volume, voxel dimensions $1 \times 1 \times 1$ mm^3), followed by BOLD fMRI (2D gradient-echo EPI, 30/2,000 ms TE/TR, 70° FA, 64 × 64 matrix, 30 slices per volume, voxel dimensions $3.125 \times 3.125 \times 5$ mm^3). The imaging data were acquired as part of a larger test battery: subjects received a 15-min orientation session in an MRI simulator, and then performed two task runs in the scanner, separated by ~10 min of other behavioral tests. For this analysis, only data from the second run were used, in order to avoid systematic error due to learning effects, seen in behavioural performance in the first run. Task stimuli (Trails A/B) were contrasted against Baseline stimulus, discarding two transition scans at the start of each block, which gave 64 scans total, per subject.

Data Processing and Analysis

BOLD EPI data were primarily preprocessed with AFNI utilities [Cox, 1996] in the following order. The EPI data were corrected for physiological effects using RETROICOR, which models and removes periodic signals correlated with externally-measured cardiac and respiratory phases [Glover et al., 2001]. Afterward, slice-timing correction was applied, using *3dTshift* with Fourier interpolation. Rigid-body motion correction was then performed with *3dvolreg*, to register all volumes to the 10th scan-volume of Run 2, for each subject. This motion correction method uses a weighted, linearized least-squares cost function and Fourier volume interpolation. In-plane spatial smoothing was applied with a 6.0 mm FWHM Gaussian kernel, via the *3dmerge* utility. Temporal Legendre polynomial detrending was then performed with *3dDetrend*, and subject motion covariates were regressed out through the use of motion parameter estimates (MPEs) produced by *3dvolreg* registration. To avoid overfitting and multicollinearity, PCA was performed on the six MPE time-courses, and the two largest-variance principal components used as regressors [Woods et al., 1998]. For all subjects, these two components accounted for more than 85% of temporal motion-parameter variance, and were most strongly influenced by pitch and inferior-superior motion parameters (see Fig. 6). Brain masks were generated using the FSL Brain Extraction Tool [Smith et al., 2004], and individual-subject SPMs transformed into a common space with the flirt registration utility [Jenkinson and Smith, 2001]. Spatial normalization of SPMs was obtained by aligning subject T1s using an iterative process, similar to the method of Guimond et al. [2000]: (1) register T1 volumes to MNI 152 (Montréal Neurological Institute template); (2) average the masked, registered T1 brain volumes; and (3) re-register T1 volumes to the normalized average. Steps (2–3) were repeated three times to generate a group template with good inter-subject alignment. We performed within-subject pipeline optimization using untransformed EPI data, and the resulting post-analysis SPMs were then spatially registered to the MNI-based template, to compare activation patterns between subjects.

Individual subject analyses were performed with a PDA model, in which a PCA decomposition is applied to the preprocessed data, followed by a two-class canonical variates analysis [Mardia et al., 1979; Strother et al., 2002, 2010]. For the analyses, one class consisted of scans acquired during Task stimulus, the other of scans acquired during Baseline stimulus. Briefly, the set of P-voxel scan volumes are represented as points in P-dimensional space, on which two PCAs are performed. The PCAs are required both to reduce data dimensionality and for data denoising; the optimal set of PC dimensions is selected for the second PCA, using NPAIRS metrics, as outlined by Strother et al. [2002, 2010]. After selecting a reduced PC basis-space, the vector was computed that, for within-class covariance matrix $\boldsymbol{W}$ and between-class covariance matrix $\boldsymbol{B}$, maximizes $\boldsymbol{W}^{-1}\boldsymbol{B}$ (i.e., the first eigenvector of matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$). Transformed back into voxel-space, this vector produces a spatial brain map that maximally discriminates between activation and baseline volume classes.

Pipeline effects were measured using metrics from the NPAIRS framework. For each subject dataset, scans from the second run were split into two groups: *Split 1* consisted of the first 32-scan epoch (Trails A and B, and two Baseline blocks, discarding two transition scans from each block), and Split 2 consisted of the second 32-scan epoch. Using two-class PDA applied to the second PCAs of each split group, Z-scored Statistical Parametric Maps (SPMs) were generated for each data split-half. Reproducibility of the activation patterns (R) was estimated by treating the two split-half SPMs as P-element vectors, and computing their Pearson correlation coefficient. Predictive accuracy of the model (P) was computed by using results of Split 1 to predict the class (Task or Baseline) for each scan in Split 2 via posterior Bayesian probability, and vice-versa. The median rate of correct prediction across both splits was subsequently reported. Finally, a reproducible, Z-scored SPM (rSPM(z)) of activations was obtained: the two split-half SPMs provide a pair of Z-scores for each voxel, which can be in turn expressed as a scatterplot. These points are projected onto the first principal component axis of the scatterplot (i.e., the signal axis), and normalized by the standard deviation along the second, orthogonal principal component axis (i.e., the noise axis). For this model, R ranges from 0 to 1 and P ranges from 0.5 to 1, with perfect performance at ($R = 1$, $P = 1$). These results were computed for PDA as a function of the size of the second PC basis set, and the number of dimensions was chosen for each preprocessing pipeline using R and P, as outlined in the Evaluating Pipeline Performance section. The signal detection advantages over other PCA dimensionality estimation approaches (e.g., Bayes Information Criterion, Bayes Evidence) are discussed in Yourganov et al. [in press]. Briefly, reproducibility-based metrics are better able to capture intrinsic changes in data dimensionality caused by changes in SNR and connectivity of brain networks, for a Linear Discriminant analysis model.

Pipeline Optimization

To test the effects of preprocessing techniques, the performance of different fixed pipelines was evaluated across all subjects. First, common preprocessing effects on (R,P) metrics were identified, using Friedman nonparametric rank-tests [Conover, 1999]. It was then verified whether these effects correlate with significant common changes in SPM activation patterns, using DISTATIS, a three-way metric multi-dimensional scaling technique [Abdi et al., 2009]. The underlying between-subject heterogeneity of pipeline effects was also tested by choosing pipelines that maximize performance for each subject, again using (R, P) metrics. It was then tested whether subjects with relatively low versus high head motion estimates had a significant interaction with pipeline performance. Finally, the effect of pipeline optimization method was examined for between-subject similarity of SPMs, as well as the structure of the common SPM activation patterns, with the commonly-used similarity metrics of SPM correlation and activation overlap.

Evaluating pipeline performance

The interactions of preprocessing choices were measured via (R,P) metrics, for the fixed sequence of preprocessing

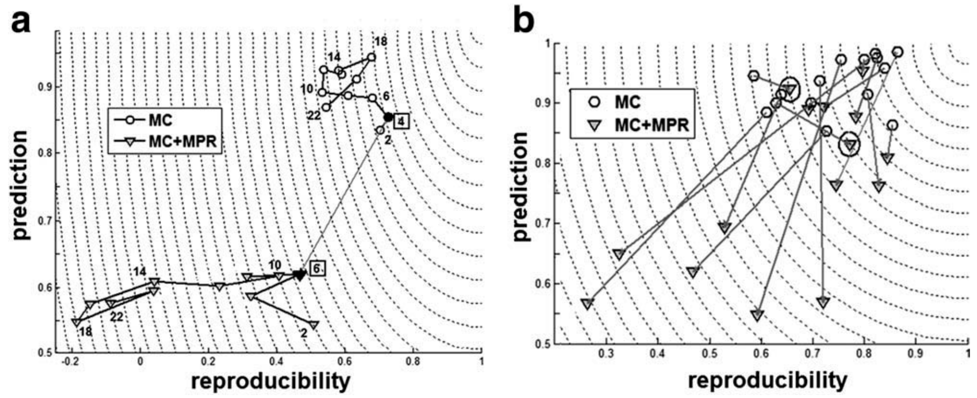


Figure 1.

Individual-subject NPAIRS analysis is performed for a range of subspaces of dimension *K*, defined by the first *K* principal components (PCs), each with reproducibility (*R*) and prediction accuracy (*P*). (**a**) A single-subject (*R*, *P*) plot is shown for two different pipeline sets, with *K* labeled for selected PCs: motion correction only (MC) and with addition of motion parameter regression (MC+MPR). The PC basis with shortest distance from (1,1) is selected for each pipeline (marked in black); dashed lines give contours of constant distance from (1,1). In this case, the best pipeline with MC,MPR performs worse than for MC-only. (**b**) The (*R*,*P*) of optimal *K* is plotted for all subjects, and both pipelines. The distribution of (*R*,*P*) is heterogeneous in both cases. It is generally worse with the addition of MPR, although two subjects are improved by adding it (circled in black).

steps: (1) physiological noise correction (PNC), (2) slice-timing correction, (3) motion correction (MC), (4) spatial smoothing, (5) motion parameter regression (MPR), and (6) temporal detrending (DET). Under this ordering, all possible combinations of PNC, MC and MPR were included/not included in the pipeline, and DET was performed for polynomial orders ranging from 0 to 5 (six different DET models). This produced $2^3 \times 6 = 48$ different preprocessing combinations; rSPM(z)s and (*R*,*P*) measures were generated for each subject across all pipelines. In the absence of evidence that one metric is more important for optimal results, pipeline performance was measured by the Euclidean distance D from the perfect model ($R = 1$, $P = 1$), which weights R and P equally. Better pipeline performance is indicated by smaller D, with $D = 0$ indicating infinite model SNR and perfect prediction. For each subject and pipeline, PDA was performed over a range of 2–22 PCs. Figure 1a shows a single-subject example of (*R*,*P*) as a function of PC number, for two different pipelines. Given that the number of PCs may significantly affect model (*R*,*P*), the optimal PC number was chosen to minimize D, for each pipeline. In Figure 1a, optimal PCs are marked in black, and the substantial change in (*R*,*P*) space between the two pipeline optima is also plotted.

Collective pipeline optimization

The (*R*,*P*) distributions across subjects are known to be, in general, heterogeneous [Shaw et al., 2003; Zhang et al., 2009]. Figure 1b plots (*R*,*P*) values for MC and MC+MPR pipelines of all 15 subjects, and shows variability in magnitude and direction of effect, with the addition of MPR. To identify consistent between-subject trends, a non-parametric measure of relative pipeline performance was used; the technique (illustrated in Fig. 2) is described as follows. For each subject, (i) obtain all pipeline *D*-values; (ii) rank them 1–48, with higher rank indicating better pipeline performance (and smaller *D*). This yields 15 sets of rankings for each fixed pipeline, obtained across all subjects. (iii) Take the median-rank (out of 15) of each pipeline, to produce a rank-normalized profile of relative pipeline performance. This profile was tested for significance using the Friedman test for multiple-treatments [Conover, 1999]; if a significant ordering was found across subjects, the fixed pipeline with highest median-rank was confirmed as optimal. A set of pipelines was also classified as not significantly different from the optimal pipeline, by performing the Nemenyi critical-difference test, at $\alpha = 0.05$ [Conover, 1999]. This is a relatively conservative test of significant differences, which uses a multiple comparison correction for all possible pairwise comparisons.

SPM effects of collective optimization

Performance-metric results were compared to the effects of preprocessing on rSPM(*z*) spatial patterns, using the DISTATIS metric multidimensional scaling technique, developed by Abdi et al. [2007, 2009] that can provide confidence interval estimates. The method displays patterns of SPM difference common across all subjects, thereby providing a technique to verify if changes in performance observed in the NPAIRS median-rank profile reflect systematic changes in activation pattern. The method is summarized as follows:

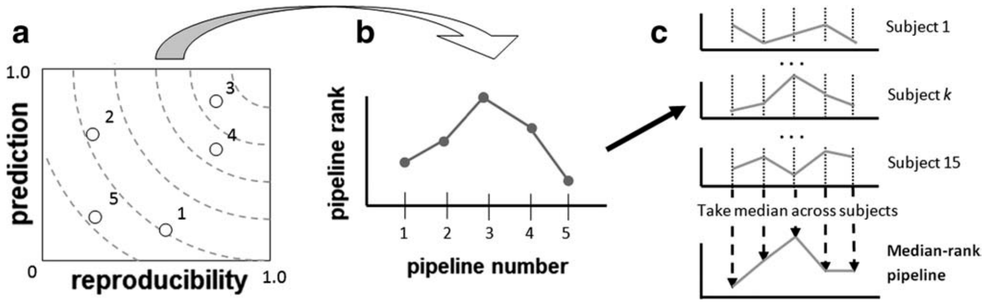


Figure 2.

Steps in identifying the optimal set of fixed pipelines across all subjects. (**a**) For subject *k*, obtain prediction and reproducibility measures for all pipelines. (**b**) For the same subject, rank pipelines by the Euclidean distance from optimal (reproducibility = 1, prediction = 1), with highest rank indicating the best performance. (**c**) Obtain pipeline ranking for all 15 subjects, then take the median rank of each pipeline, across all subjects. This produces the median-rank profile, which can be assessed for significance, using a Friedman test.

1. For the *k*th subject ($1 < k < 15$), generate the row and column-centered 48 × 48 matrix $\boldsymbol{S}_{\mathrm{k}}$ of rSPM(z) reproducibility between all possible pipeline pairs. The matrix $\boldsymbol{S}_{\mathrm{k}}$ is a doubly-centered correlation matrix, of all possible pipeline rSPM(z) pairings.
2. Form the 15 × 15 matrix $\boldsymbol{C}$ of pair-wise similarity between the 15 $\boldsymbol{S}_{\mathrm{k}}$ matrices, based on the $R_{\mathbf{V}}$ coefficient which is a measure of shared information between positive semi-definite matrices, similar to the squared correlation coefficient [Abdi et al., 2007; Robert and Escoufier, 1976].
3. Compute the eigendecomposition of matrix $\boldsymbol{C}$. The first eigenvector gives a set of coefficients α_k that weight how similar each $\boldsymbol{S}_{\mathrm{k}}$ (and thus each subject *k*) is to the strongest common $\boldsymbol{S}$-matrix pattern. Now compute this denoised, weighted compromise matrix $S_+ = \sum_{i=1}^{15} \alpha_i S_i$. The compromise matrix depicts the most common 48 × 48 pipeline pattern of cross-correlation and thus rSPM(*z*) reproducibility, across all subjects.
4. Compute the eigendecomposition of matrix $\boldsymbol{S}_+$ (equivalent to a PCA of $\boldsymbol{S}_+$), and project $\boldsymbol{S}_+$ into the new basis space. This gives factor scores for the 48 pipelines methods.

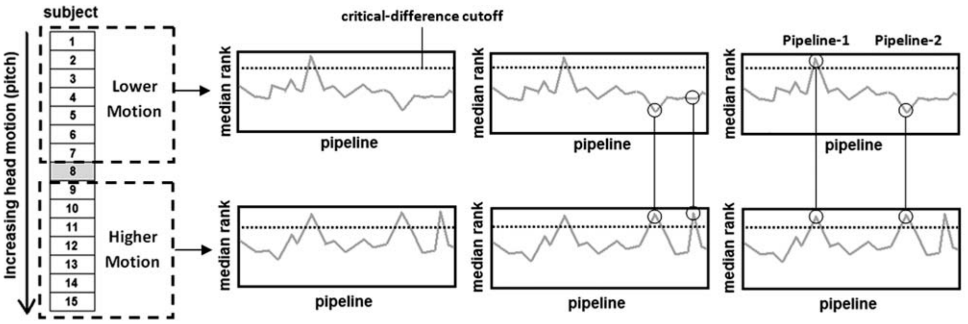
Using this method, the 48 pipeline rSPM(z)s are represented as points in 47-dimensional PCA space. The distance between any two pipelines' points is a measure of Euclidean "distance" between rSPM(z), with points that are closer together indicating more similar activation maps. Distances in this space may be converted to correlation, for a measure of reproducibility between pipelines. Confidence estimates were also produced for groups of pipelines, by generating bootstrap samples of the correlation matrices (1,000 iterations). These matrices were projected into the same principal component space, and the resultant 95% confidence ellipses were drawn. The directions of maximum (*R*,*P*) increase in DISTATIS space were also plotted, by computing the correlations between *R* and *P* metrics and PC-scores of the SPMs in each dimension, rescaled by [1/2] of the square root of each dimension's eigenvalue; for a more extensive discussion of projecting supplementary variables into PC-space, see Abdi et al. [2007, 2009].

Individual pipeline optimization

The between-subject heterogeneity of pipeline effects was also tested using (*R*,*P*) measures, along with the performance effects of individually optimizing pipelines. The specific set of preprocessing choices that gives the minimum *D*-value was found for each subject, and labelled the individually-optimized pipeline set. Trends were examined in the individually-optimized set of preprocessing steps, and (*R*,*P*) values of individually-optimized pipeline were compared to the optimal fixed pipeline.

Subject motion effects

Based on an expected interaction of head motion with preprocessing steps, the flexibility of fixed-pipeline techniques was demonstrated by measuring the influence of head-motion range on preprocessing performance; Figure 3 summarizes the steps in this analysis. For a given subject, six rigid-body MPEs were used to estimate the displacement of each scan-volume relative to a reference scan. To characterize the variability of within-run head motion, temporal standard deviation was measured for each MPE, computed across all scan-volumes in the run. This provides a summary estimate of the relative amount of within-run head-motion range on all six movement axes, for a given subject. The "pitch" axis of motion (nodding movement), which generally demonstrated the largest standard deviation of subject head movement, was used to test for pipeline interactions. Figure 3a: subjects were sorted into high and low-motion groups, ranking them based on pitch standard deviation and discarding the median-ranked



a Order subjects into high- and low-motion groups, based on pitch-axis standard deviation.

b Obtain plots of median rank as a function of pipeline, for each motion group

c Identify pipelines that are optimal for one group and not the other, based on the critical-difference cutoffs.

d Identify SPM effects of going from best whole-group pipeline (**ppln-1**) to the one most affected by head motion (**ppln-2**), for each.

Figure 3.

Schematic of steps in identifying interaction of head motion and preprocessing pipelines. For (**b**), median-rank profiles and the critical-difference boundary from the optimal pipeline (at $p = 0.05$) are computed as in Collective Pipeline Optimization. For (**c**), pipelines different between the two groups are tested for significance using a permutation test on ranks. For (**d**), Pipeline-1 is the optimal median-rank profile of the whole group (*Collective Pipeline Optimization*), while Pipeline-2 has the greatest difference in median rank between the two groups.

subject, which provided seven subjects per group. Figure 3b: for each of the motion-range groups, pipelines were ranked 1-48 based on D, and the median-rank profile of D was estimated across the 7 subjects (as in Collective Pipeline Optimization), with a Nemenyi critical-difference boundary at $p = 0.05$. Figure 3c: preprocessing pipelines were identified, if their median-rank was within the critical-difference boundary for one motion group, but outside the boundary in the other group, as (R,P) values of these pipelines are affected by range of head motion. Each pipeline in this subset was tested, to confirm significant difference due to head motion, by performing a permutation test on ranks in the two groups. For a given pipeline of interest, high- and low-motion subject labels were thus exchanged under all possible permutations, and the difference in sum of ranks calculated between high- and low-motion groups for each permutation. From these values, an empirical probability distribution of mean-differences was obtained, which was used to estimate the probability of the difference in the original high- vs. low-motion classification occurring at random. Permutation-test probability values were thresholded at a False-Discovery Rate (FDR) of 0.05 [Genovese et al., 2001], to adjust for multiple tests.

Figure 3d: two pipelines of interest were then selected, to explore the effects of motion range on preprocessing in greater detail. The first, Pipeline-1, has optimal fixed-pipeline performance, identified in Collective Pipeline Optimization, and it has the highest median-rank for all subjects. The second, Pipeline-2, demonstrates greatest difference in rank between the two motion groups. The fraction of activated rSPM(z) voxels for Pipeline-2 relative to Pipeline-1 was measured, at a threshold of FDR = 0.05. This was performed separately for both high and low-motion groups, to compare the differential effects of Pipeline-2 on extent of activation. For each motion-group, the mean rSPM(z) was also computed across subjects, for Pipelines-1 and -2, and the activation maps compared, again at threshold FDR = 0.05. This procedure allows the comparison of spatial differences in activation by pipeline, for the two motion groups. The mean rSPM(z)s were computed across subjects, to estimate the most reliable spatial signal for each group. As first established by Cochran [1937], the weighted combination of independent, normal variables (e.g., Z-scored subject voxel-values) that minimizes variance and maximizes reliability is the average of these variables, after normalization by their respective standard deviations. In this case, this is the voxel-wise mean, computed over individual Z-scored activation maps.

Group SPM effects

The effects of pipeline optimization method on between-subject SPM heterogeneity were examined, for four specific optimization schemes. For both the (F) fixed-pipeline and (I) individually-optimized pipeline sets, a 15 × 15 matrix of pairwise correlation was computed between subject SPMs, termed the Between-Subject Reproducibility (BR). Taking the matrix row (or column) median produces 15 estimates of subject similarity relative to the rest of the group; this set of values is the BR distribution for pipeline sets 1 and 2. Because these two pipeline sets are based on optimization performed on an individual-subject basis, for comparison, two additional pipeline sets were identified that directly maximized the similarity of between-subject SPMs. The third pipeline set (B) was chosen to give the highest attainable median BR, and the fourth set, (A) maximized another commonly-used similarity metric, activation overlap (AO) [Rombouts et al., 1997]. This measure is

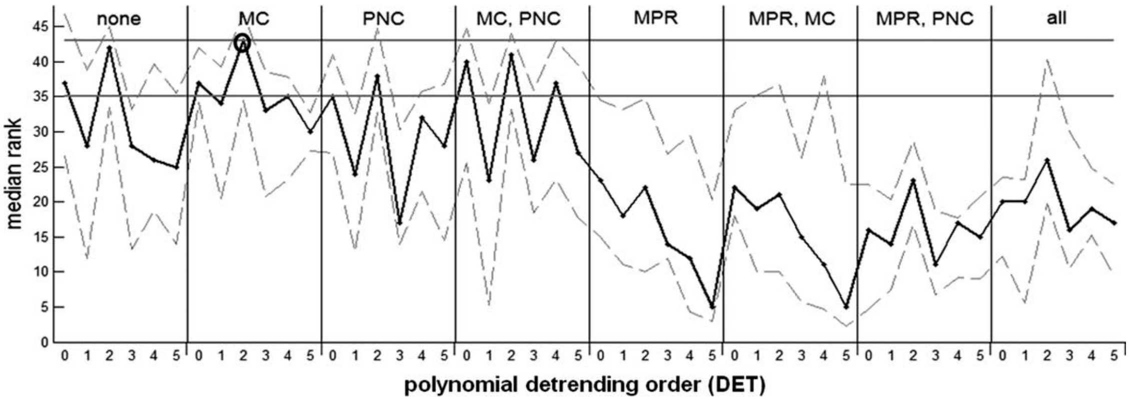


Figure 4.

A profile of relative pipeline performance, based on the distance *D* from perfect prediction and reproducibility. Pipelines are ranked for each subject, and the median rank is taken across subjects, for each pipeline. Tested pipeline steps include Motion Correction (MC), Physiological Noise Correction (PNC), Motion Parameter Regression (MPR) and temporal detrending (DET) with Legendre polynomial orders 0–5. All possible combinations of preprocessing steps were tested under a fixed order (PNC,MC,MPR), for 48 pipelines in total. The black curve gives median rank, and dashed curves indicate upper and lower distribution quartiles. The optimal pipeline of MC,DET2 is given by highest rank-median; all pipelines within grey horizontal critical-difference lines are not significantly different from the optimum (Nemenyi critical difference test at $p = 0.05$).

based on the Jaccard statistic: for two SPMs X and Y, voxels are declared active if the value is above some critical threshold c, and overlap is thus expressed as:

$$\mathrm{AO}(X,Y) = \left|\frac{X \cap Y}{X \cup Y}\right|_{|x|,|y|>c}$$

Activation was defined as all voxels above the threshold FDR = 0.05, computed separately for each SPM, for which the Jaccard metric gives a fraction of shared overlap, ranging from 0 to 1. Along with BR measures of these four pipeline sets, the per-voxel standard deviation of rSPM(z) values was computed across all subjects, which generated four standard deviation brain maps. These maps display the spatial distribution of relative between-subject signal variability under the four optimization schemes.

The effect of optimization on the group activation structure was also examined. For the four pipeline sets, the distribution of between-subject AO was generated, again based on a 15 × 15 pairwise comparison matrix. This estimates the congruency between significant regions of activation, across all subjects. For a spatial representation of the effects of optimization, the rSPM(z)s were also averaged across all subjects, for each of the four pipeline sets. As discussed in Subject Motion Effects, averaging rSPM(z)s provides the most reliable estimate of Z-scored spatial activation across all subjects. For all absolute Z-score thresholds greater than 1.0, consistent differences in the number of both activated and deactivated voxels were seen as a function of optimization method. Representative plots are shown at a fixed Z-score of 3.0, to demonstrate the common trends in these effects.

RESULTS

Collective Subject Optimization

The median-ranked performance of each pipeline is plotted in Figure 4, based on distance D from the ideal ($R = 1$, $P = 1$). Higher median rank indicates smaller D and better performance across subjects for a given preprocessing set. A significant fixed-pipeline effect was found across all subjects ($p < 0.001$, Friedman test). The highest median-rank occurs for a relatively low-processing pipeline of MC,DET2. However, there is a subset of seven other pipelines that are also not significantly worse, as they lie within the $p = 0.05$ critical-difference boundary. Detrending is one of the most consistent determinants of optimal performance, as all pipelines optimize with DET0 or DET2 (with the latter generally being higher-ranked). The minimum-processing pipeline of just DET0 is not significantly worse than MC,DET2, but it must be emphasized that this occurs after the PCA denoising inherent to PDA analysis. This pipeline also has extreme variance in ranks (more variable than MC,DET2, $p = 0.04$ by nonparametric median-centered Ansari-Bradley test), making it an unsuitable preprocessing method. The detrending orders DET(0,2,4) also generally outperform DET(1,3,5), indicating that the order (e.g., odd or even) of the highest polynomial basis significantly affects the optimal (R,P) of analyses. This ordering effect is consistent with the correlation between the task design and the polynomial bases: the task-design vector, convolved with a standard HRF function, exhibits correlation with Legendre polynomials of orders 1–5, of [−0.186, −0.018, −0.270, −0.077, −0.184], respectively. Orders 1,3,5 show significant negative correlation ($p < 0.05$) whereas orders 2,4 are non-significantly correlated.

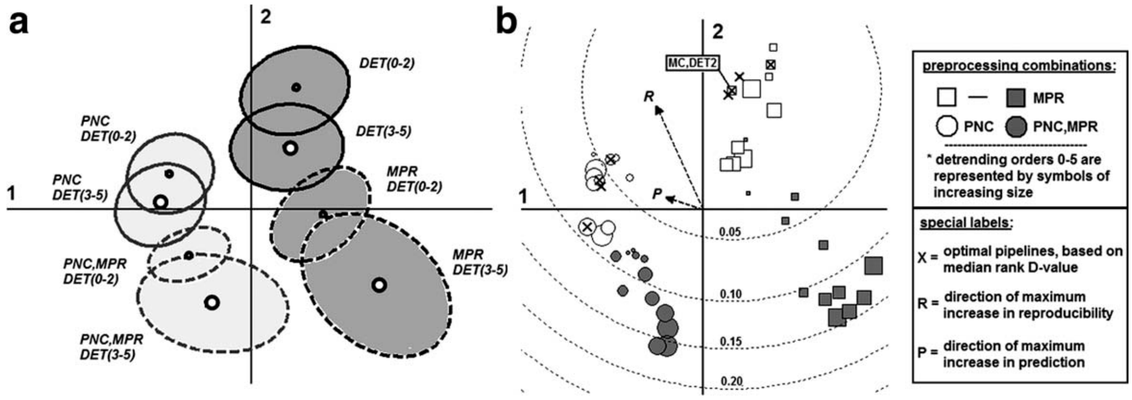


Figure 5.

The differences in NPAIRS statistical parametric maps (SPMs) by pre-processing pipeline, that are most common across all subjects, are plotted in principal component dimensions 1 and 2 of a multidimensional scaling analysis (28% total variance). SPM similarities are based on pairwise Pearson correlation. Pipelines with physiological noise correction (PNC) and motion parameter regression (MPR) are identified, along with detrending order (DET). (**a**) 95% confidence ellipses are given for the centroids of pipelines with the four different combinations of PNC and MPR, and high or low detrending (orders 0–2 and 3–5, respectively). Variance is given by bootstrap estimates of cluster-mean locations, based on 1,000 iterations. (**b**) Individual SPMs are shown in principal component space, along with directions of maximum increase in reproducibility (*R*) and prediction (*P*). The 8 optimal pipelines based on fixed-pipeline ranking (see Fig. 4) are marked with an **X**: DET(0,2), MC,DET(0,2), PNC,DET2 and PNC,MC,DET(0,2,4). Curves of constant correlation difference, relative to the fixed-pipeline optimum MC,DET2, are given by the dashed contour lines, with correlation difference from the optimal SPM labeled on contours.

Regarding other preprocessing components, MPR-based pipelines have a consistent negative effect on *D*, leading to significantly lower ranks in median performance. Pipelines with MPR perform worse with increasing detrending order, but the addition of PNC tends to both reduce between-subject variability, and alter detrending effects. PNC also interacts directly with detrending order, as adding this step to pipeline DET0 causes significantly worse performance, whereas adding it to MC,DET4 significantly improves rank (crossing the critical-difference boundary). MC generally increases median rank, as rank tends to increase with its addition to any preprocessing set. However, its median-rank effect tends to depend on detrending order, with only marginal improvement at optimal orders DET0,2.

The median-rank profile, if based on either *R* or *P* metric alone, remains similar to *D*. Pipeline rankings under *R* and *P* alone have (mean ± standard deviation) Spearman correlations with *D* of 0.86 ± 0.12 and 0.74 ± 0.20 across all subjects ($p < 0.001$, all subjects). The metrics also optimize at similar pipelines as *D*, although the highest median-rank *R* occurs at a more preprocessed pipeline of MC, PNC,DET2, whereas median-rank *P* maximizes at a less preprocessed set of MC,DET0.

SPM Effects of Collective Optimization

The effects of pipeline choice on spatial activation maps were also verified, using DISTATIS. Figure 5a shows the relative differences between pipeline-generated rSPM(z)s, common across all subjects in DISTATIS-space representation, for Dimensions 1 and 2 (15 and 13% of total variance). Mean 95% confidence ellipses are plotted for of all four combinations of MPR and PNC included/excluded, under low-order detrending (DET0-2 combined) and high-order detrending (DET3-5 combined), giving confidence estimates of eight different centroids. For DISTATIS analysis, the greatest SPM difference is due to increasing detrending order, and not between odd/even DET orders, as observed in the *D*-value rank profile of Figure 4. Pipelines with/without MC are not distinguished, as no common MC effect on SPMs occurs in this basis space. Although not shown here, significant MC effects are expressed in dimensions 3 and 4 (9 and 7% total variance), along with a DET interaction.

Changes across dimension 1 are driven (from right to left) by the addition of PNC, and more weakly by the removal of MPR (as ±MPR ellipses overlap along Dim. 1). Dimension 2 shows an effect (from top to bottom) of increasing detrending order and addition of MPR; DET and MPR induce similar effects, causing a progressive change in rSPM(z) pattern. For example, there is a 95% confidence overlap between pipelines with DET3-5, and those with MPR,DET0-2. However, the addition of PNC (right to left, Dim. 1) not only causes a significant, common change in activation structure for all pipelines, but this addition also tends to decrease the heterogeneity of both DET and MPR effects, as the SPM-group centroids with PNC included cluster together more tightly, with smaller confidence ellipses.

In Figure 5b, individual SPMs are plotted in this DISTATIS space. Pipeline SPMs with increasing DET-order are

TABLE I. Optimal pipeline choice is shown for each subject, with the change in reproducibility (*R*), prediction (*P*), and *D*-metric, relative to the optimal fixed pipeline of motion correction and second-order temporal detrending

Subject-optimal pipeline				Metric changes			Absolute performance			
MPR	PNC	MC	DET	Δ*R*	Δ*P*	Δ*D*	PC#	R	*P*	*D*
		X	2	0	0	0	2	0.855	0.863	0.200
X		X	0	0.005	0.003	−0.006	16	0.644	0.918	0.365
	X	X	2	0.011	−0.012	−0.008	4	0.836	0.962	0.169
	X		2	0.016	−0.027	−0.010	6	0.880	0.958	0.127
			2	0.011	−0.001	−0.011	18	0.811	0.972	0.191
		X	0	0.022	0.013	−0.025	18	0.862	0.971	0.141
		X	1	0.041	−0.001	−0.041	20	0.862	0.982	0.140
			2	0.083	−0.017	−0.059	2	0.809	0.836	0.251
	X	X	0	0.062	0.0052	−0.061	10	0.760	0.905	0.258
	X	X	0	0.062	0.025	−0.067	2	0.868	0.940	0.146
	X	X	5	0.101	−0.003	−0.096	14	0.814	0.934	0.198
X		X	2	0.143	−0.068	−0.101	14	0.771	0.831	0.284
		X	0	0.102	0.013	−0.103	16	0.856	0.985	0.144
	X		1	0.124	0.074	−0.138	20	0.735	0.958	0.268
			5	0.199	−0.037	−0.184	18	0.784	0.909	0.235

Subjects highlighted in grey are optimized at pipelines identified as optimal via group-wise testing, as shown in Figure 4.

depicted with increasing symbol size, pipelines with MPR included are shown as darker symbols, and pipelines with/without PNC are represented by circles and squares, respectively. The dashed contours of constant correlation difference, relative to fixed optimum MC,DET2, show that adding PNC weakly reduces rSPM(z) similarity to mean cluster correlation of 0.93 ± 0.01, and changes in detrending order have a weaker, variable effect. Maximum-processing pipelines of PNC,MPR,DET3-5 give the lowest mean correlation of 0.86 ± 0.03. The directions in which *R* and *P* increase as a function of preprocessing pipeline are also projected into the DISTATIS space. Reproducibility is most strongly correlated with SPM changes in DISTATIS-space, improving in the direction of low-order DET and pipelines without MPR, indicating that these steps most strongly influence *R*. Prediction is driven predominantly by Dim. 1 (PNC and MPR effect), though less so than reproducibility; prediction weakly increases with both removal of MPR and addition of PNC. The best-performing pipelines, defined by the rank-profile of Figure 4 (e.g. above the critical-difference line), are also marked by "**X**" in Figure 5b. Although the optimal pipelines are non-significantly different with the inclusion/exclusion of PNC, as measured by median rank, the addition of PNC gives a significantly different SPM activation pattern. Consistently different activation patterns may thus possess comparable levels of (*R*,*P*) demonstrating the additional information that may be obtained with the DISTATIS procedure.

Individual Subject Optimization

The results of individual subject optimization are given in Table I, where the pipeline choice maximizing *D* for each subject is listed, along with the change in performance from the optimal fixed pipeline MC,DET2. Reproducibility is consistently improved for 14/15 subjects, with mean Δ*R* of 0.065 ± 0.059 (a mean change of 10.4% ± 9.9% from fixed-pipeline *R*), whereas prediction shows a weak, variable effect, with mean Δ*P* of 0.003 ± 0.031 (a mean change of −0.2% ± 3.4% from fixed-pipeline *P*); optimization consistently tends to increase the within-run homogeneity of SPMs. Note that 9/15 subjects, highlighted in grey, have pipelines included in the set of eight optimal fixed pipelines (e.g., within the critical difference bounds of Fig. 4), with mean optimization Δ*D* of 0.038 ± 0.035. The other 6/15 subjects are optimized with pipeline sets that otherwise tend to perform very poorly. For example, two subjects optimize with MPR, which has significant negative effect on group performance, as noted in the previous sections. Optimizing the 6-subject group with non-typical optimal pipelines gives mean Δ*D* of −0.094 ± 0.064, with a larger effect than the group of nine with standard pipeline optima ($p = 0.06$, unpaired Wilcoxon test). The 9/15 subjects with pipelines not significantly different from the fixed-optimum MC, DET2 are optimal across a wide range of PCs, although only 3/9 subjects require more than 10 PCs for optimal performance. The second group deviates extensively in optimal preprocessing set, and consistently optimizes at larger subspace sizes of 14–20 PCs. The number of subjects optimized with each preprocessing pipeline is summarized in Table II. MC is generally an optimal step to apply for 10/15 subjects, while PNC has a relatively heterogeneous effect; MPR remains generally detrimental under individual subject optimization, optimizing for 2/15 subjects. A fixed detrending order is also generally detrimental, as all orders cause suboptimal performance for the majority of subjects. However, higher detrending orders DET3-5 are generally ineffective,

TABLE II. This table shows the fraction of subjects with each pre-processing step included in their optimal pipelines

	Pre-processing step added			Order of temporal detrending polynomial					
	PNC	MC	MPR	DET(0)	DET(1)	DET(2)	DET(3)	DET(4)	DET(5)
Fraction optimized	6/15	10/15	2/15	5/15	2/15	6/15	0/15	0/15	2/15

Preprocessing steps are abbreviated as: motion correction (MC), physiological noise correction (PNC), motion parameter regression (MPR) and polynomial detrending (DET).

with only two subjects optimized in this range. DET0 and DET2 evidenced the greatest frequencies of optimization at 5/15 and 6/15 subjects, respectively, which is similar to fixed-pipeline effects (Collective Subject Optimization).

Subject Motion Effects

Figure 6 presents the distribution of within-run standard deviation across all subjects, for each of the six rigid-body MPEs. Despite complex motor responses during the experiment, head motion effects were generally low, with only two cases of mean uncentered pitch (nodding movement) exceeding 1.0°. Furthermore, the standard deviations of the motion estimates are consistently below 0.3 mm/°, with a single exception for pitch movement at 0.59°. Subjects were separated into high- and low-motion groups based on standard deviation of pitch, with respective group averages of 0.28° ± 0.16° and 0.07° ± 0.03°.

After splitting subjects into high- and low-motion groups, median-rank profiles of pipeline performance were obtained for the high and low-motion groups (see Fig. 7); both groups show significant fixed-pipeline ordering effect ($p < 0.001$, Friedman test). Eleven different pipelines were found within the critical-difference boundary for one subject group and not the other, indicating a head-motion interaction with the pipeline. For 2/11 pipelines, significantly different median ranks were measured between groups (circled in Fig. 7, bottom), based on a permutation test of group differences and corrected for multiple comparisons at FDR = 0.05. PNC, DET0 is significantly worse for high-motion subjects, while MPR,PNC,MC,DET2 performs significantly better for high-motion subjects. There is no consistent effect of adding MC as a function of head motion, as pipelines with MC are not consistently different in median rank between the motion groups. For the lower-motion set (Fig. 7, top), all pipelines without MPR are generally high-ranked, with most such pipelines at or near the confidence boundary; the pattern is somewhat similar to that of Figure 4 for all subjects. Higher-motion subjects (Fig. 7, bottom) show more significant detrending effects across all pipeline combinations, with more variance in rank across subjects as well, shown by wider interquartile ranges, particularly under addition of MPR. Nevertheless, the previously-identified group optimum of MC,DET2 remains the best median pipeline for both groups—this optimum is thus stable, independently of motion range.

In Figure 8a the boxplots show fractional change in the number of active voxels at FDR = 0.05, when PNC and MPR are added to the fixed-pipeline optimum MC, DET2, for both high- and low-motion groups. There is a significant difference in median effect, wherein the lower-motion group has a median of 50% of active voxels removed; higher-motion subjects have a median increase in active voxels of 21%, but with a wider distribution of values among subjects. The spatial map effects of adding PNC and MPR to the group optimum MC, DET2 are shown in Figure 8b; the mean rSPM(z)s are compared between the two pipelines, at threshold FDR = 0.05. For the low-motion case, there is a reduction in extent of the large cerebellar, occipital and parietal activations, and smaller dorsolateral frontal and medial clusters are removed entirely. For higher motion, although right medial-temporal clusters are reduced, the spatial extent of posterior activation increases, and left prefrontal activations also appear. These results indicate a differential interaction between subject motion range and preprocessing choices, for nominally active voxels.

Group SPM Effects

Distributions of BR are shown in Figure 9a, for the four different optimization schemes. The first two methods do

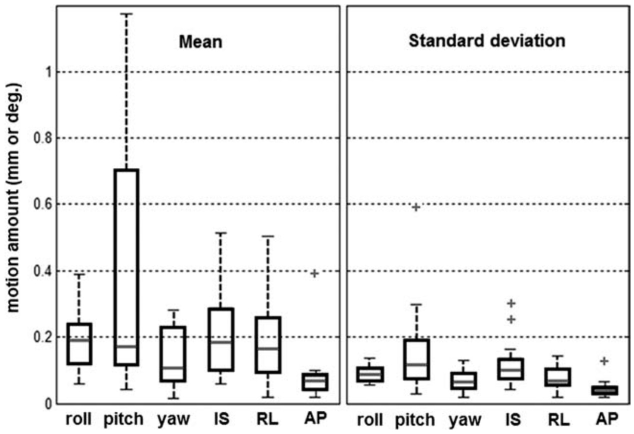


Figure 6.

Distribution of median rigid-body motion parameter estimates, across all subjects, based on motion correction output. "Mean" data gives average absolute brain-volume displacement relative to the reference volume, across the whole run. "Standard deviation" is also given for brain-volume displacement. Rotation axes include roll, pitch, and yaw, and translation axes include inferior-superior (IS), right-left (RL) and anterior-posterior (AP).

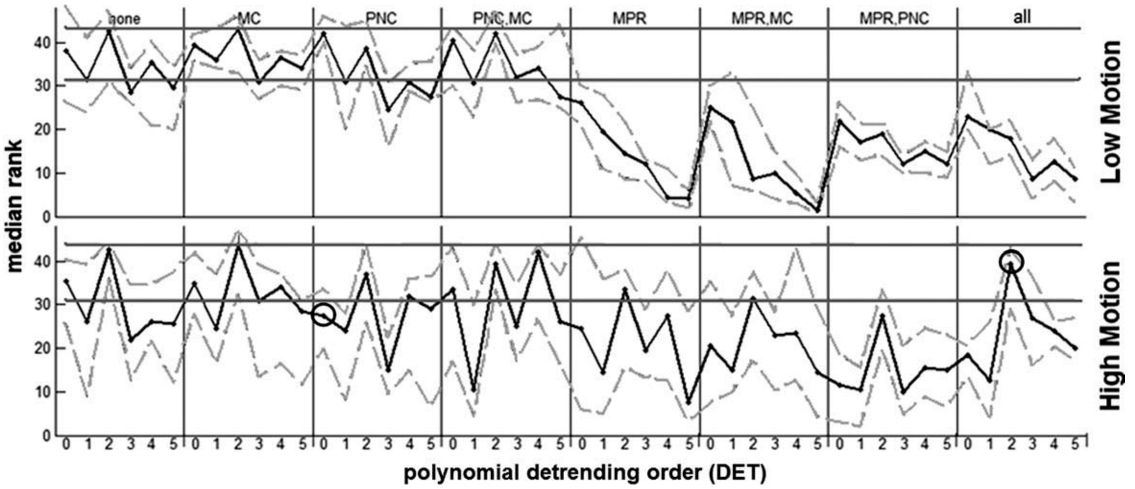


Figure 7.

Median rank profiles of pipeline performance for low and high-motion subject groups, classified based on median split of estimated "pitch" head movement; this represents (Steps b–c) in Figure 3. (*Top*) low-motion subjects, significance $p < 0.001$ (Friedman test). (*Bottom*) high-motion subjects, significance $p = 0.003$ (Friedman test). Pipelines circled in black show significant high-motion difference in ranks from low-motion, under permutation testing, at FDR threshold 0.05. The optimal pipeline is given by highest median rank, and lower horizontal grey lines indicate the critical difference boundary from the optimum ($\alpha = 0.05$).

not incorporate between-subject SPM similarity: (F) the fixed pipeline MC,DET2, which minimizes group-median *D*, and (I) the pipeline set individually chosen to minimize *D* for each subject. The latter two maximize between-subject SPM similarity, based on (B) the median BR value, and (A) the median AO value. The pipeline set maximizing BR were generally similar to those minimizing *D* (Table III); the only consistent change was an increase in the fraction of subjects optimizing with higher order detrending DET3-5, and MPR at marginal significance of $p = 0.11$ (by nonparametric sign-test). Comparing (F) fixed-pipeline and (I) individual optimization, there is no significant change in median BR, but the spread of BR values has increased in the latter case, demonstrating specific incidences of both increased and decreased pairwise SPM similarity between subjects. Note that SPMs are intrinsically heterogeneous across subjects, given a maximum median BR of 0.45 for pipeline set (B). The AO-maximizing

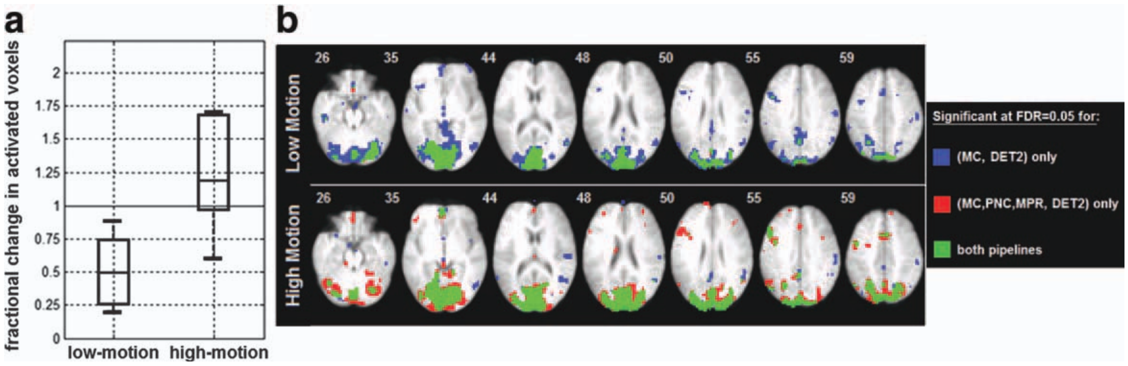


Figure 8.

Effects of adding physiological noise correction and motion parameter regression (PNC,MPR) to the optimal fixed-pipeline of motion correction and second-order detrending (MC,DET2). Results are shown for high and low-motion subject groups, based on a median split of subject "pitch" movement (Step d in Fig. 3). (**a**) The fractional change in number of activated voxels across subjects (at FDR = 0.05) is shown, for low- and high-motion subject groups. (**b**) Changes in the mean activation map for both groups, with addition of (PNC,MPR), at FDR = 0.05. Blue regions indicate activation only present with MC,DET2, red regions are activations only present when MPR,PNC are added, and green patterns show activation that are significant for both pipelines choices (e.g., irrespective of whether PNC,MPR are added). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

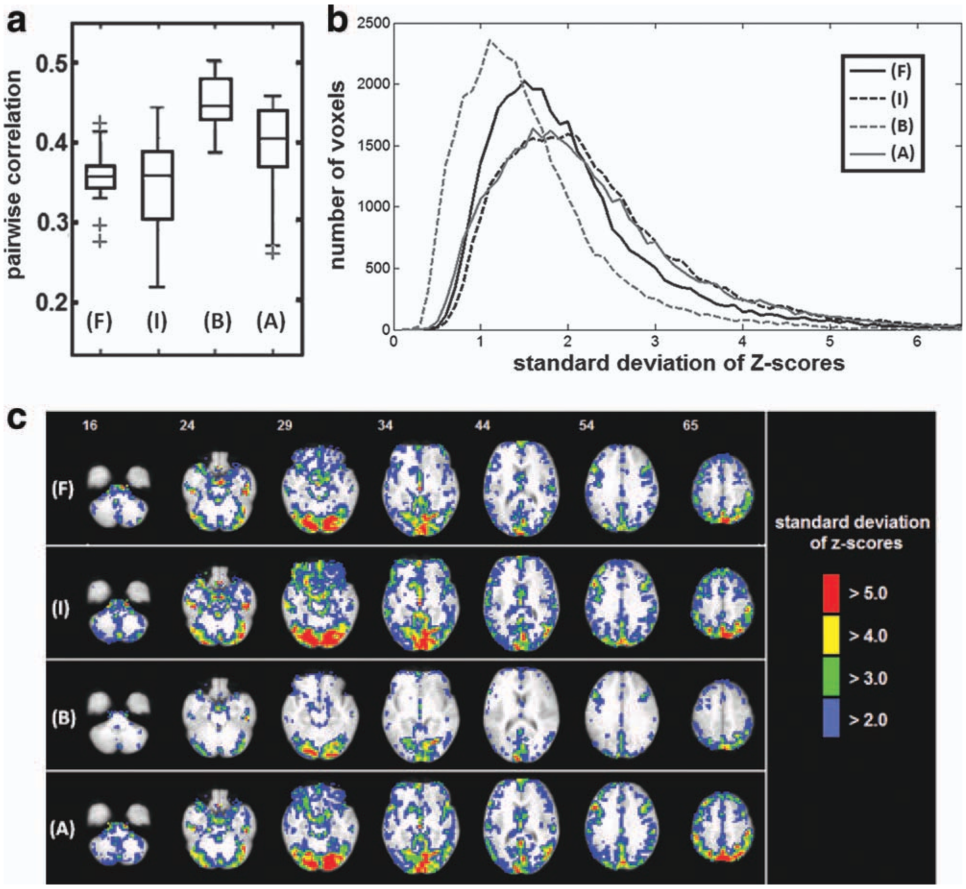


Figure 9.

Effects of pipeline choice on between-subject reproducibility and variance are shown, for (F) the optimal fixed-pipeline of motion correction and second-order detrending, (I) individual-subject pipeline optimization, (B) the pipeline set maximizing between-subject reproducibility, (A) the pipeline set maximizing activation overlap between subject SPMs. (**a**) The distribution of between-subject reproducibility, based on pairwise activation map correlation, for the four optimization methods. For each voxel, *Z*-score standard deviation is also computed across all subjects, for each optimization method. (**b**) The histogram of between-subject voxel standard deviation is shown, for each pipeline optimization method. (**c**) Spatial maps of between-subject standard deviation in *Z*-scores are shown for each pipeline optimization method. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

pipeline set (A) also produces relatively high BR values. Figure 9b plots the histogram of standard deviation in voxel *Z*-scores across subjects, for the four pipeline optimization schemes. The individually-optimized (I) and AO-maximizing (A) pipelines have similar distributions of higher variance, while the BR-maximizing pipeline (B) has the lowest peak in variance values; the fixed-pipeline optimum (F) has an intermediate distribution of variance between (I,A) and (B). Figure 9c shows spatial maps of voxelwise standard deviation in rSPM(z) signal, across subjects, for the 4 optimization methods. If compared with respective mean *Z*-scored activation maps for each method (Fig. 10c) it can be seen that variance is generally maximal at the regions of strongest task activation. This further demonstrates relatively high levels of between-subject heterogeneity, localized to regions of high *Z*-score signal, which are minimized if the BR-maximizing pipeline set is chosen.

Plots of between-subject AO are provided in Figure 10a. As with BR, the set of pipelines maximizing AO is similar to those minimizing D (Table III), although MPR and PNC became consistently more important at $p = 0.063$ and $p = 0.14$, respectively (by nonparametric sign-test), and higher-order detrending DET3-5 tends to be more important for optimization. Individual-subject optimization (I) demonstrates significantly increased shared overlap relative to the fixed pipeline (F). This indicates that individual subject optimization extracts a maximized region of strong common activation, but also causes generally higher between-subject variance in Z-scores in this region (as

TABLE III. This table demonstrates the fraction of subjects with each preprocessing step in their pipeline, when subject pipelines are selected to jointly maximize Activation Overlap (AO), at False-Discovery Rate threshold FDR = 0.05, and Between-Subject Reproducibility (BR)

	Preprocessing step added			Order of temporal detrending polynomial					
Optimized metric	PNC	MC	MPR	DET(0)	DET(1)	DET(2)	DET(3)	DET(4)	DET(5)
BR	5/15	11/15	6/15	4/15	2/15	5/15	2/15	1/15	1/15
AO	10/15	12/15	7/15	3/15	1/15	5/15	1/15	2/15	3/15

shown in Fig. 9b). The relatively low activation overlap for BR-maximizing SPMs (B), combined with generally low voxel-wise variance across subjects, indicates that this method is strongly affected by sub-threshold (e.g., low *Z*-score) signal structure. Figure 10b also shows the number of activated (top) and deactivated (bottom) voxels for absolute *Z*-score threshold 3.0; as discussed in Group SPM Effects, these results are representative of trends in signal for all *Z*-scores whose magnitude is larger than 1.0. Individual subject optimization (I) provides both the second-highest extent of task activation, and the highest range of task deactivation of all tested pipeline sets. The BR-maximizing pipeline (B) also has the lowest extent of both activation and deactivation structures.

Figure 10c displays the *Z*-scored SPM for each pipeline set, averaged across all subjects. All pipelines demonstrate significant positive task activation, localized to upper cerebellar, occipital and parietal regions, along with more sparse dorsolateral prefrontal activations. Furthermore, the four pipelines all produce negative activations in left temporal-parietal, posterior cingulate and ventromedial prefrontal regions which are all associated with the so-called "Default-Mode Network." This network is increasingly active during cases of "mind wandering" when a task (in this case, the baseline task) has lower cognitive engagement [Greicius et al., 2004]. The individually-optimized pipeline set (I) shows increased deactivation strength and spatial extent, particularly in the dorsolateral region, as compared with other optimization methods. The pipeline set maximizing group AO (A) gives the greatest extent and strength of task activation, particularly emphasizing bilateral cerebellar activation and more extensive prefrontal activation; it also produces the lowest deactivation signal strength and spatial extent.

DISCUSSION

We have presented a flexible general framework for evaluating preprocessing choices and their interactions in individual subjects, and applied it to a range of temporal preprocessing choices. Applying novel nonparametric testing procedures to NPAIRS performance metrics, and adapting the DISTATIS multidimensional scaling methods to evaluate preprocessing pipelines, we show that this methodology provides useful complementary information on the effects of pipeline optimization. We have also established the general heterogeneity of subject response to preprocessing, and applied all of these techniques to present the first evaluation of interactions between standard motion and physiological noise correction methods.

The presented analysis techniques are particularly useful for testing general pipeline effects due to their robustness to outliers. On the individual-subject level, *R* and *P* metrics tend to identify pipeline choices that minimize outliers across repeated epochs, and the rSPM(Z) computation minimizes outlier influence on activation maps. When comparing results across subjects, the nonparametric Friedman test methods are robust to subject variability, due to both rank-normalization and the conservative Nemenyi test for significant effects. In addition, DISTATIS is specifically designed to downweight outlier subjects, and extract the most common pattern of SPM similarity across subjects, with bootstrap confidence estimates quantifying the significance of these results. The increased median and reduced spread of BR values for individually-optimized preprocessing (Fig. 10a, I), compared with fixed pipelines (Fig. 10a, F), demonstrates that these robust techniques produce an associated increase in the homogeneity of SPM spatial overlap between subjects.

Using our nonparametric testing framework, significant effects of fixed pipelines applied across all subjects were identified for motion and physiological noise correction techniques. In addition, replicating previous findings, we demonstrate that the choice of temporal detrending polynomial order is a dominant effect for fixed preprocessing pipelines. It is known that detrending method has a significant impact on fMRI results, in both univariate [Tanabe et al., 2002] and multivariate frameworks [LaConte et al., 2005; Shaw et al., 2003]. Fixed-pipeline performance is also dependant on whether highest detrending order is even or odd, with the former offering better (*R*, *P*). This effect may be due to the intrinsic frequency of the task design, which is best approximated by a 7th-order polynomial. The odd polynomial bases, which are more correlated with task design, are likely regressing out components of task-associated signal. Motion correction, which has previously been found to improve performance significantly in individual-subject PDA analyses [Zhang et al., 2009], also shows a small but consistent effect on fixed-pipeline performance, and generally improves performance for individual-subject optimization; the relatively weak effect may be due to the low subject head motion in our data set, despite the complex motor movement task. This limited head motion may be a

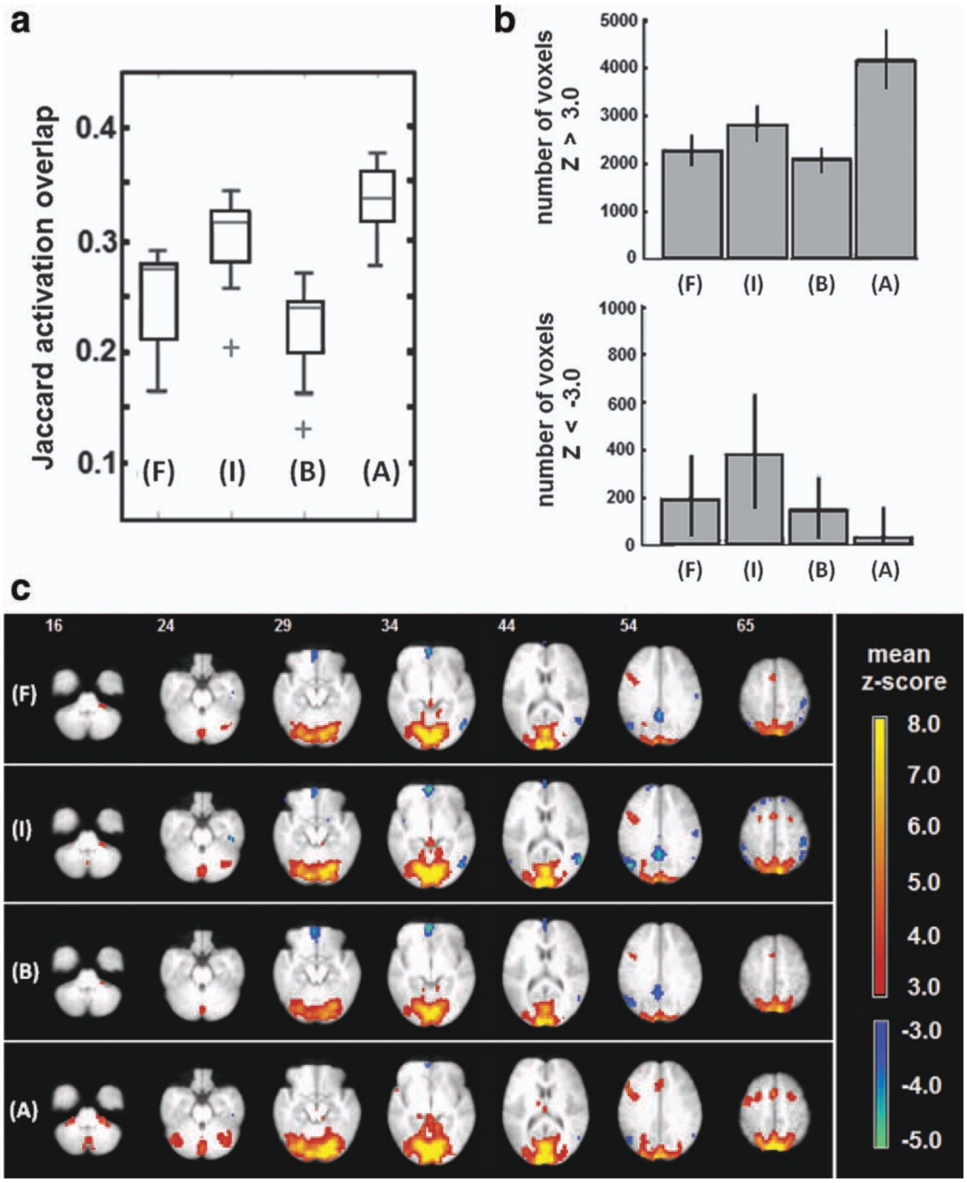


Figure 10.

Effects of pipeline choice on between-subject activation overlap and mean signal are shown, for (F) the optimal fixed-pipeline of motion correction and second-order detrending, (I) individual-subject pipeline optimization, (B) the pipeline set maximizing between-subject reproducibility, (A) the pipeline set maximizing activation overlap between subject SPMs. (**a**) Distribution of between-subject activation overlap, based on pairwise Jaccard overlap (at FDR = 0.05), for the four optimization methods. The mean rSPM(z) is also computed across all subjects, for each optimization method. (**b**) Number of activated (top) and deactivated (bottom) voxels exceeding absolute *Z*-score of 3.0 are shown for the averaged rSPM(z)s from all four pipeline sets. Error bars are estimated based on a bootstrap of the average rSPM(z)s (1,000 iterations). (**c**) Spatial maps of mean rSPM(z)s, for optimized pipeline sets. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

consequence of our careful attempts to train subjects to remain still, prior to fMRI data acquisition.

While our results show a significant negative effect when applying MPR, a previous PDA-based study by Evans et al. [in press] has shown that MPR improves group reproducibility. The possible causes of this discrepancy seem to relate to differences in the amounts of subject motion between the two studies, and are discussed below. Our findings also demonstrate highly heterogeneous PNC subject effects for (*R*,*P*), which demonstrate that, along with detrending, this is one of the more important steps to optimize when selecting subject pipelines. It is possible that inconsistent PNC effects are due to sub-optimal ordering of preprocessing steps, which Jones et al. [2008] have shown to have weak but significant effect on the temporal variance of EPI data. Our own preliminary investigations [Churchill et al., 2010] have also shown that the ordering of PNC, particularly relative to slice-timing correction, significantly affects the activation patterns in resultant SPMs.

Using DISTATIS, we also established that preprocessing choice induces strong, consistent effects on SPM activation pattern, and that these effects reflect the significant differences measured using the median-ranks based on optimized NPAIRS (*R*,*P*) metrics. Our DISTATIS analysis confirmed the dominant effects of DET and MPR, and demonstrated that they induce similar spatial effects; this may be due to the similar low-order regression designs of the two techniques, as it has been observed that motion induces low-order, near-linear drift effects [Liu et al., 2001; Lund et al., 2005]. Interestingly, weaker MC effects were expressed primarily in DISTATIS dimensions 3 and 4, which suggest that the PC denoising step in PDA analysis may effectively correct for motion artifact; MC effects are orthogonal to PNC and MPR, which is unsurprising given that MC is the only technique that is not regression-based, and is thus different from the other two. In addition, DISTATIS results were insensitive to even/odd detrending order, unlike the NPAIRS-based rank profile. This possibly due to the insensitivity of DISTATIS to temporal patterns in fMRI data, suggesting that the rSPM(Z) patterns are generally robust to this order effect. These findings indicate the potential additional information provided by multivariate analysis of metric patterns over the selection of single optimal metrics per subject [see also LaConte et al., 2003]. We suggest that a combined analysis of metric patterns of SPM similarity and NPAIRS performance metrics yields information that cannot be extracted using any other single technique.

Significant intra- and inter-subject heterogeneities were also found across all preprocessing steps, reflected in both the variability of median-rank profiles, and the significant pipeline variability under individual subject optimization. It was found that the individual-subject pipeline minimizing *D* consistently improves subject within-run reproducibility relative to the optimal fixed pipeline, whereas prediction shows comparatively weak effects (Individual Subject Optimization). The dominance of reproducibility when minimizing *D* suggests that prediction is near-optimal at fixed pipeline MC,DET2 for this contrast, whereas reproducibility is more consistently sub-optimal at this fixed-pipeline optimum. It was also noted that under pipeline optimization, negative default-mode activation regions are consistently identified as part of the discriminant pattern for this visuo-motor task. This demonstrates that DMN-like networks may be extracted using standard preprocessing methods, without requiring global normalization [Chang et al., 2009; Murphy et al., 2009], which has been found to introduce spurious task-negative activations, or ICA-based analysis methods [Beckmann et al., 2005].

Our findings also indicate that relatively high (*R*,*P*) may be obtained with pipeline optimization for comparatively small datasets acquired with only a few minutes of scanning, rather than the more conventional acquisition period for fMRI neurobehavioural tests (>1 h). This confirms our expectations that (*R*,*P*) are dependent on intrinsic data SNR, which is controlled by preprocessing. For stationary signal/noise (and Gaussian noise distribution), a larger sample size would both increase *R* and improve the stability of (*R*,*P*) estimates. However, if changes in cognitive and physiological responses occur over time, longer acquisition times may introduce increased signal variability, reducing experimental power and (*R*,*P*) estimates. The effects of sample size are therefore likely to be strongly experiment and subject dependent. How to choose pipelines for (*R*,*P*) optimization across even larger sample sets is an issue that warrants additional investigation in the future.

Subjects optimize for a relatively wide range of pipelines, with the split between subjects that perform best for a group-optimal pipeline ("conforming" group), and those with significantly different pipeline optima ("non-conforming" group) at 60–40%, as established in Individual Subject Optimization. This demonstrates extensive heterogeneity of pipeline effects even in a young, healthy subject set, perhaps linked to the increased signal variability that may be associated with accurate task performance [McIntosh et al., 2008]. Although the high significance of the fixed-pipeline ranking method suggests that subjects have generally similar pipeline responses, our results indicate that varying pipelines on an individual-subject basis may still significantly enhance individual task-contrast results, and measures of group similarity. This effect may be particularly strong in small-sample data sets, where even a few strong artifacts can significantly influence PCA data decomposition, and thus PDA results, requiring specialized preprocessing; we are currently extending our results to a range of analysis models with and without initial PCA decomposition. However, the common trend in which "non-conforming" subjects consistently optimize at higher PC counts than those required for "conforming" subjects suggests an alternate explanation. It has been shown that weaker activation signals in PDA require more PCs to maximize performance [Evans et al., in press; Strother et al., 2004; Yourganov et al., in press]. As SNR drops, task structure becomes more distributed across the principal component spectrum. Our results demonstrate that subjects with weaker, more distributed signal, relative to noise (as indicated by the optimal PC dimensionalities), may require specialized (e.g., non-typical) pipeline choices to optimize activation maps.

An interaction was identified between pitch-based head movement and preprocessing effects, despite relatively low levels of subject head-motion. Subjects with increased head motion evidence a clear benefit from performing MPR, along with PNC. Ollinger et al. [2009] showed that MPR removes noise and signal uncorrelated with motion, along with movement artifacts, and that MPR tends to remove a greater extent of task activation under strong BOLD haemodynamic response. We suggest that because of the intensity-based cost function for registration, MC is affected by the largest sources of signal variance. If motion variance is sufficiently low that its signal change in raw fMRI data becomes smaller than that of other sources, including task-effect and physiological noise, MPR will start to remove signal correlated with these latter variance sources. Because of the strong task activation and low head motion in our experimental data, it is thus expected that MPR has a generally detrimental effect on activation. This hypothesis is consistent with the results of results of Evans et al. (in press), in which MPR significantly improves the performance of children with much higher levels of motion and more distributed BOLD responses. Our results also expand on findings of Freire and Mangin [2001], who recommend applying MPR to fMRI data with lower subject motion: however, there may be a task- and signal-dependant "lower limit" of motion, where the MPR step becomes deleterious.

The comparison of SPM structure between subjects also provided insight into the effect of pipeline choice on group similarity and the extracted activations. The differential effect of fixed vs. individually optimized pipelines is demonstrated by different pairwise similarity measures. Individual subject optimization significantly increases the spatial extent of shared activation relative to the fixed-pipelines, but it offers no improvement in median pairwise correlation. This increased heterogeneity of between-subject reproducibility is expected, given that individual subject brain responses are inherently variable, potentially more so for young, healthy participants [McIntosh et al., 2008]. Furthermore, our results indicate that stable group responses are at least partly a result of using fixed preprocessing pipelines, and that inter-subject differences, along with subject-dependent pipeline optimization, need to be more carefully considered than is usual in the current literature [Miller et al., 2009]. Individual subject optimization increases between-subject signal variance relative to the fixed-pipeline set, particularly for nominally task-activated regions with higher mean Z-score. In addition, we have strong evidence that the increased intersubject variability seen for individually optimized versus fixed pipelines is not an artifact of applying heterogeneous pipelines. The pipeline sets minimizing within-subject D (Fig. 9I) and those maximizing between-subject AO (Fig. 9A) both increase intersubject variance relative to the fixed pipeline set; however, the heterogeneous BR-maximizing pipelines significantly reduce intersubject variance. This demonstrates that non-fixed preprocessing choices across subjects may either increase or reduce heterogeneity of subject SPMs, depending on the optimization criterion. Furthermore, we emphasize that the increased intersubject variability and decreased BR values of individually-optimized preprocessing, shown in Figure 9a(I), is driven by increased variability in regions of significant activation. This is clearly illustrated by the increased between-subject activation overlap at FDR = 0.05 (Fig. 10a(I)) with individual subject optimization.

In generating the mean Z-scored SPMs, we have also shown that choice of pipeline may strongly influence the type of detected brain networks. Similar results have been previously shown for PDA in group and individual subject analyses [Chen et al., 2006; Strother et al., 2004], where optimizing based on either prediction or reproducibility generated significant differences in specific activation regions with different responses to signal artifacts. Our results directly show the advantage of individually-optimizing subject pipelines based on D metric: this method both maximizes within-run reproducibility for each subject, and provides the second-highest overlap task activation, for the tested optimization methods; it also maximizes the strength and extent of otherwise weaker and more transient default-mode regions (Fig. 9, Method 2). It thus provides a compromise among the tested pipeline optimization methods, sensitive to the greatest overall range of brain response both within- and between-subjects, but with relatively high between-subject task variance.

We have also shown that the use of other performance metrics yield their own unique results, which may also be of practical interest. AO maximization emphasizes the maximal extent of task-associated networks, including otherwise absent frontal and cerebellar structures. This measure is driven by high Z-score signal due to the FDR = 0.05 threshold, and it produces maximal task-activation extent and mean Z-score compared to other optimization methods, at the expense of elevated between-subject signal variance, to which it is insensitive. On the other hand, BR maximization, while reducing activation extent (Fig. 9a, Method 3), gives a "consensus map" of the most consistently expressed structures across subjects, as evidenced by the low standard deviation in voxel-wise signal across subjects (Fig. 8b). These results also suggest that our metric of spatial reproducibility can be specifically chosen to extract brain networks of interest. Our approach may be applied on an individual subject basis, because of the flexibility of the NPAIRS framework. One may replace the correlation measure of R, for example with a measure of activation or deactivation overlap—which would preferentially maximize positive and negative brain networks, respectively (see Fig. 10).

Regardless of the metric chosen for individual subject optimization, care must also be taken to avoid confounds due to circular analysis, that is, using the optimization criterion to maximize some hypothesized effect and then using these results to confirm the prediction [Kriegeskorte et al., 2009]. Given the low CNR and sensitivity to pipeline choice in fMRI data, it is possible to reinforce artifactual signal in a brain region, which does not necessarily reflect the true BOLD response, using circular analysis methods. For example, using pipeline optimization to maximize signal in a region of interest, in order to confirm the presence of this activation, will heavily bias the estimates and therefore invalidate the conclusions. It is thus important to carefully dissociate the optimization metric from any assumptions implicit in the tested hypotheses. Our individual-subject results have been employed to (1) determine the relative effect of pipeline choice on within-subject (R,P), and (2) test the effect of pipeline optimization on inter-subject SPM variability. Within-subject optimization effects are presumably independent of between-subject reproducibility, thus avoiding the issues of circular analysis.

This experiment has presented a general framework for assessing preprocessing pipelines, along with the effects of pipeline optimization on a subject-wise basis. It also provides a novel examination of the differential brain response between task activation and baseline conditions of the Trails-Making Task, for young healthy participants. The current preprocessing findings are, as such, measures specific to the tested stimulus contrast and low subject motion data of this group. The heterogeneity of pipeline effect among the subjects indicates a likely increase in the importance of optimization for cohorts that are younger and older, and/or ill. These groups move more in the scanner [Seto et al., 2001], and the strength and spatial structures of artifacts and their task-responses are significantly more heterogeneous [Bullmore et al., 1999; Grady et al., 2006]. These groups represent our future targets for the results of this study applied to clinical-batteries in which Trails A-B is one potential task component. This framework is suited to identify common and/or uniquely optimal preprocessing effects on a task and/or subject-dependent basis. Our current results also further demonstrate that data collections in short sessions of each task (e.g., for clinical design) may readily yield significant insights into the brain's functional structure, provided a careful choice of preprocessing pipelines is made.

REFERENCES

Abdi H, Valentin D, Chollet S, Chrea C (2007): Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. Food Qual Prefer 18:627–664.

Abdi H, Dunlop JP, Williams LJ (2009): How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). NeuroImage 45:89–95.

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

Army Individual Test Battery (1944): Manual of Directions and Scoring. Washington, DC: War Department, Adjutant General's Office.

Bannister PR, Brady JM, Jenkinson M (2004): TIGER—A New Model for Spatio-Temporal Realignment of FMRI Data. Berlin, Heidelberg: Springer-Verlag.

Beckmann CF, DeLuca M, Devlin JT, Smith SM (2005): Investigations into resting-state connectivity using independent component analysis. Philos Trans R Soc Lond B Biol Sci 360:1001–1013.

Birn RM, Diamond JB, Smith MA, Bandettini PA (2006): Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. NeuroImage 31:1536–1548.

Bullmore ET, Brammer MJ, Rabe-Hesketh S (1999): Methods for diagnosis and treatment of stimulus-correlated motion in generic brain activation studies using fMRI. Hum Brain Mapp 7:38–48.

Chang C, Cunningham JP, Glover GH (2009): Influence of heart rate on the BOLD signal: The cardiac response function. Neuroimage 44:857–886.

Churchill NW, Abdi H, Strother SC (2010): The Effects of Ordering Motion Correction and Physiological Noise Correction in fMRI Analyses. Proceedings 16th Annual Meeting OHBM, Barcelona, Spain. pp 152.

Cochran WG (1937): Problems arising in the analysis of a series of similar experiments. J R Stat Soc Supp 4:102–118.

Conover WJ (1999): Practical Nonparametric Statistics, 3rd ed. Weinheim: Wiley.

Cox RW (1996): AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

Della-Maggiore V, Chau W, Peres-Neto PR, McIntosh AR (2002): An empirical comparison of SPM preprocessing parameters to the analysis of fMRI data. NeuroImage 17:19–28.

Evans JW, Todd RW, Taylor MJ, Strother SC (2010): Group specific optimization of fMRI processing steps for child and adult data. NeuroImage 50:479–490

Folstein MF, Folstein SE, McHugh PR (1975): "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12:189–198.

Freire L, Mangin JF (2001): Motion correction algorithms may create spurious brain activations in the absence of subject motion. NeuroImage 14:709–722.

Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RSJ (1995a): Spatial registration and normalization of images. Hum Brain Mapp 2:165–189.

Friston KJ, Frith CD, Frackowiak RSJ, Turner R (1995b): Characterizing dynamic brain responses with fMRI: A multivariate approach. NeuroImage 2:166–172.

Genovese CR, Lazar NA, Nichols T (2001): Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15:870–878.

Glover GH, Li TQ, Ress D (2001): Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn Reson Med 44:162–167.

Grady CL, Springer MV, Hongwanishkul D, McIntosh AR, Winocur G (2006): Age-related changes in brain activity across the adult lifespan. J Cogn Neurosci 18:227–241.

Greicius MD, Srivastava G, Reiss AL, Menon V (2004): Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. Proc Natl Acad Sci 101:4637–4642.

Guimond A, Meunier J, Thirion J (2000): Average brain models: A convergence study. Comput Vis Imag Understanding 77:192–210.

Hachinski V, Iadecola C, Petersen RC, Breteler MM, Nyenhuis DL, Black SE, Powers WJ, DeCarli C, Merino JG, Kalaria RN, Vinters HV, Holtzman DM, Rosenberg GA, Dichgans M, Marler JR, Leblanc GG (2006): National institute of neurological disorders and stroke—Canadian stroke network vascular cognitive impairment harmonization standards. Stroke 37:2220–2241.

Hu X, Le TH, Parrish T, Erhard P (1995): Retrospective estimation and correction of physiological fluctuation in functional MRI. Magn Reson Med 34:201–212.

Jiang A, Kennedy DN, Baker JR, Weisskoff RM, Tootell RBH, Woods RP, Benson RR, Kwong KK, Brady TJ, Rosen BR, Belliveau JW (1995): Motion detection and correction in functional MR imaging. Hum Brain Mapp 3:224–235.

Johnstone T, Walsh KS, Greischar LL, Alexander AL, Fox AS, Davidson RJ, Oakes TR (2006): Motion correction and the use of motion covariates in multiple-subject fMRI analysis. Hum Brain Mapp 27:779–788.

Jones TB, Bandettini PA, Birn RM (2008): Integration of motion correction and physiological noise regression in fMRI. NeuroImage 42:582–590.

Kay KN, David SV, Prenger RJ, Hansen KA, Gallant JL (2007): Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. Hum Brain Map 29:142–156.

Kim B, Boes JL, Bland PH, Chenevert TL, Meyer CR (1999): Motion correction in fMRI via registration of individual slices into an anatomical volume. Magn Reson Med 41:964–972.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009): Circular analysis in systems neuroscience: The dangers of double dipping. Nat Neurosci 12:535–540.

LaConte S, Strother S, Cherkassky V, Anderson J, Hua X (2005): Support vector machines for temporal classification of block design fMRI data. NeuroImage 26:317–329.

Liu TL, Frank LR, Wong EC, Buxton RB (2001): Detection power, estimation efficiency, and predictability in event-related fMRI. NeuroImage 13:759–773.

Lund TE, Nbrgaard ND, Rostrup E, Rowe JB, Paulson OB (2005): Motion or activity: Their role in intra- and inter-subject variation in fMRI. NeuroImage 26:960–964.

Mardia K, Kent J, Bibby J (1979): Multivariate Analysis. London, United Kingdom: Academic Press.

McIntosh AR, Kovacevic N, Itier RJ (2008): Increased brain signal variability accompanies lower behavioral variability in development. PLoS Comput Biol 4:e1000106.

Miller MB, Donovan CL, Van Horn JD, German E, Sokol-Hessner P, Wolford GL (2009): Unique and persistent individual patterns of brain activity across different memory retrieval tasks. Neuroimage 48:625–635.

Moeller JR, Strother SC (1991): A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. J Cereb Blood Flow Metab 11:A121–A135.

Morgan VL, Dawant BM, Li Y, Pickens DR (2007): Comparison of fMRI statistical software packages and strategies for analysis of images containing random and stimulus-correlated motion. Comput Med Imaging Graph 31:436–446.

Murphy K, Birn RM, Handwerker DA, Jones TB, Bandettini PA (2009): The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? Neuroimage 47:1092–1104.

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

Oldfield RC (1971): The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9:97–113.

Ollinger JM, Oakes TR, Alexander AL, Haeberli F, Dalton KM, Davidson RJ (2009): The secret life of motion covariates. NeuroImage 47:S122.

Orchard J, Atkins MS (2003): Iterating Registration and Activation Detection to Overcome Activation Bias in fMRI Motion Estimates. Berlin, Heidelberg: Springer-Verlag.

Poline JB, Strother SC, Dehaene-Lambertz G, Egan GF, Lancaster JL (2006): Motivation and synthesis of the FIAC experiment: reproducibility of fMRI results across expert analyses. Hum Brain Mapp 27:351–359.

Robert P, Escoufier Y (1976): A unifying tool for linear multivariate statistical methods: the RV-coefficient. Applied Statistics, 25:257–265.

Rombouts S, Barkhof F, Hoogenraad F, Sprenger M, Valk J, Scheltens P (1997): Test-retest analysis with functional MR of the activated area in the human visual cortex. AJNR 18:1317–1322.

Sarty GE (2007): Computing Brain Activation Maps from fMRI Time-Series Images. Cambridge University Press, Cambridge, UK.

Shaw ME, Strother SC, Gavrilescu M, Podzebenko K, Waites A, Watson J, Anderson J, Jackson G, Egan G (2003): Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics. NeuroImage 19:988–1001.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy R, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23:208–219.

Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, Frutiger S, Muley S, LaConte S, Rottenberg D (2002): The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. NeuroImage 15:747–771.

Strother SC, LaConte S, Hansen LK, Anderson J, Zhang J, Pulapura S, Rottenberg D (2004): Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. NeuroImage 23:S196–S207.

Strother SC, Oder A, Spring R, Grady C (2010): The NPAIRS Computational Statistics Framework for Data Analysis in Neuroimaging. Proc. 19th Int. Conf. Computational Statistics, Paris, France.

Stuss DT, Bisschop SM, Alexander MP, Levine B, Katz D, Izukawa D (2001): The trails making test: A study in focal lesion patients. Psychol Assess 13:230–239.

Tam F, Churchill N, Strother S, Graham S (2010): System for computerized writing and drawing during fMRI. Proc Intl Soc Mag Reson Med, 17th Scientific Meeting, p 1704.

Tanabe J, Miller D, Tregellas J, Freedman R, Meyer FG (2002): Comparison of detrending methods for optimal fMRI preprocessing. NeuroImage 15:902–907.

Thomas CG, Harshman RA, Menon RS (2002): Noise reduction in BOLD-based fMRI using component analysis. Neuroimage 17:1521–1537.

Woods RP, Grafton ST, Cherry SR Mazziotta JC (1998): Automated image registration: I. General methods and intrasubject, intramodality validation. J Comput Assisted Tomogr 22:139–152.

Yourganov G, Chen X, Lukic A, Grady C, Small S, Wernick M, Strother SC: Dimensionality estimation for optimal detection of functional networks in BOLD fMRI data. Neuroimage (in press).

Zhang J, Liang L, Anderson JR, Gatewood L, Rottenberg DA, Strother SC (2008): A java-based fmri processing pipeline evaluation system for assessment of univariate general linear model and multivariate canonical variate analysis-based pipelines. Neuroinformatics 6:123–134.

Zhang J, Anderson JR, Liang L, Pulapura SK, Gatewood L, Rottenberg DA, Strother SC (2009): Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second- level CVA. Magn Reson Imaging 27:264–278.