ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 33 (2006) 471–481

Reducing interscanner variability of activation in a multicenter fMRI study: Controlling for signal-to-fluctuation-noise-ratio (SFNR) differences

Lee Friedman,[a,*] Gary H. Glover,[b] and The FBIRN Consortium[c,1]

[a]Department of Psychiatry and Human Behavior, University of California-Irvine, Irvine, CA 92617, USA
[b]Department of Radiology, Stanford University, Stanford, CA 94305, USA
[c]FBIRN—Functional Testbed-Biomedical Informatics Research Network, NIH-NCRR, Bethesda, MD 87131, USA

Received 23 April 2006; revised 6 July 2006; accepted 20 July 2006
Available online 6 September 2006

Variation in scanner performance will lead to variation in activation patterns in multicenter fMRI studies. The purpose of this investigation was to evaluate the effect of statistically covarying for scanner differences in signal-to-fluctuation-noise-ratio (SFNR) on reducing scanner differences in activation effect size as part of a multicenter fMRI project (FIRST BIRN). For SFNR, "signal" is typically the mean intensity over time and "fluctuation noise" is the temporal standard deviation. Five subjects were sent to 9 centers (10 scanners) and scanned on two consecutive days using a sensorimotor fMRI protocol. High-field (4 T and 3 T) and low-field (1.5 T) scanners from three vendors (GE, Siemens and Picker) were included. The effect size for the detection of neural activation during a sensorimotor task was evaluated as the percent of temporal variance accounted for by our model (percent of variance accounted for, or PVAF). Marked scanner effects were noted for both PVAF as well as SFNR. After covariate adjustment with one of several measures of SFNR, there were dramatic reductions in scanner-to-scanner variations in activation effect size. Variance components analyses revealed 75%–81% reductions in variance due to scanner with this method. Thus, controlling for scanner variation in SFNR may be an effective method to homogenize activation effect sizes in multicenter studies.

Introduction

Multicenter studies are becoming increasingly common in MRI research because they provide several advantages over unicenter studies: the capability to study rare diseases by increasing the catchment area for recruitment (Casato et al., 2005; Bevan et al., 2002), the capability to accumulate large samples for situations in which the effect size under study is small or where it is very important to be extremely certain about the presence of an effect, the capability to test drug effectiveness in multiple localities and contexts (O'Connor et al., 2004; Brada et al., 2001) and the capability to support complex high-order statistical modeling (e.g., structural equation modeling). Of particular current interest is the capability of doing large scale genetic association studies relating various allelic patterns to brain structure and function (Insel et al., 2004). Aside from the issue of undersampling rare genetic patterns, Freimer and Sabatti (2003) have highlighted the need for "… comprehensive assemblages of systematically collected phenotypic information…" and proposed The Human Phenome Project. The inclusion of human brain structure and function in such a project will likely require multicenter MRI studies. Most such multicenter research has involved structural MRI (Ewers et al., 2005; Schnack et al., 2004), but multicenter studies based on fMRI (Casey et al., 1998; Stocker et al., 2005; Zou et al., 2005; Friedman and Glover, 2006; Friedman et al., 2006) and other modalities (Chang et al., 2004; Silver et al., 1999) are beginning to appear.

Multicenter studies inevitably involve differences in hardware and software across member sites, especially when more than one vendor's instrumentation is employed. These differences can lead to systematic, site-dependent effects in fMRI sensitivity (Zou et al., 2005; Friedman et al., 2006). Thus, when attempting to pool fMRI data across centers, it is important to understand and control for confounding site effects in order that generalizable conclusions can be drawn about the population(s) being studied.

One approach to reducing variation from one MRI center to another is to only include centers with exactly the same scanner hardware, although this may unduly limit some of the advantages of multicenter studies noted above. Another approach, adopted by the FBIRN project (http://www.nbirn.net/TestBeds/Function/index.htm), is to develop methods to reduce scanner-induced variations in imaging characteristics when multiple types of scanners are employed. Interscanner differences in activation patterns will add noise to such multicenter fMRI investigations. One can simply

* Corresponding author. 1312 Michael Hughes Dr. NE, Albuquerque, NM 87112, USA.
E-mail address: lfriedman10@comcast.net (L. Friedman).
[1] www.nbirn.net.
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2006.07.012

treat scanner variance as noise, one can try to correct for scanner differences by modifying the imaging data prior to analysis (Friedman et al., 2006, Thomason et al., 2006) or one can try to adjust for scanner effects statistically. In the present study, the latter approach is employed.

Analysis of covariance (ANCOVA) is a particularly appropriate tool for statistical adjustment of scanner effects. Let us take the case of testing for a difference between Group A and Group B in a multicenter study. Proper experimental design would require that reasonable samples of both groups be acquired at all sites. In general, and except for initial testing, the subjects at Center x cannot be conveniently studied at any other center. ANCOVA in this case is being used as a statistical matching procedure (Tabachnick and Fidell, 1989), to adjust group means to what they would be if all subjects scored identically (at the overall mean level) on the chosen covariate. The test comparing Group A to Group B is thus performed as if all subjects were collected at one imaginary average center. In choosing a covariate, Tabachnick and Fidell (1989) point out that "…the goal is to identify a small set of covariates that are uncorrelated with each other but correlated with the dependent variable. Conceptually, one wants to select covariates that adjust the dependent variable for predictable but unwanted sources of variability".

Scanners differ in several ways that will affect brain activation patterns (Friedman and Glover, 2006), including field strength, signal to noise ratio, stability, ghosting levels, drift, T2*-weighting, etc… In the present study, we have chosen to focus on scanner performance differences in a particular type of signal-to-noise ratio (SNR), i.e., signal-to-fluctuation-noise-ratio or SFNR (Glover and Lai, 1998). This is a measure of SNR that applies to functional studies in which multiple brain volumes are collected consecutively over time and is defined on a voxel-wise basis as the mean intensity (signal) divided by the temporal standard deviation (fluctuation noise). It is well documented that scanners differ on SFNR when studied with phantoms (Friedman and Glover, 2006) and with humans (Kruger and Glover, 2001; Kruger et al., 2001; Triantafyllou et al., 2005). Furthermore, it is well documented that manipulations that enhance SFNR also enhance activation effect size (Lowe and Sorenson, 1997; Parrish et al., 2000). Therefore, SFNR appears to be an excellent choice as a covariate.

In the present report, we analyzed fMRI data gathered during the "FBIRN Phase I" study, in which the same 5 subjects were scanned on 10 scanners around the United States. Subjects participated in 2 fMRI scanning visits on separate days. The results on the sensorimotor (SM) task (a task which produced particularly robust activation) were analyzed before and after adjustment for several forms of SFNR. We closely examine the effects of this adjustment on the "scanner effect", assessed in two ways.

Materials and methods

Subjects

Five healthy, English-speaking males (mean age: 25.2, range = 20.2 to 29) participated in this study. All were right-handed, had no history of psychiatric or neurological illnesses and had normal hearing in both ears. Each subject traveled to 9 sites (10 scanners) (Table 1), where they were scanned twice over a period of 2 days for a total of 20 scans per participant. There were no missing subject visits (scan sessions), that is, all 100 visits (5 subjects × 2 visits × 10 scanners) were available for analysis. However, for 1 subject, data from one visit were excluded for technical reasons related to the data acquisition. Therefore, the analyses presented herein are based on 99 subject visits. All subjects were instructed to avoid alcohol the night before the study, caffeine 2 h prior to the study and to get a normal night's sleep the night before a scan session. Informed consent was obtained from every subject before participation in this study and before every scan session. The approval of every site's Internal Review Board was obtained before conducting this study.

Image acquisition

A bite bar was used to stabilize each subject's head and was placed in the subject's mouth and affixed to the coil at the beginning of each scan session. An initial T2-weighted, anatomical volume for functional overlay was acquired for each subject (fast spin-echo, turbo factor = 12 or 13, orientation: parallel to the AC–PC line, number of slices = 35, slice thickness = 4 mm, no gap, TR = 4000 ms, TE = approximately 68, FOV = 22 cm, matrix = 256 × 192, voxel dimensions = 0.86 mm × 0.86 mm × 4 mm). The parameters for this T2 overlay scan were allowed to vary slightly from scanner to scanner according to field strength or other local technical factors. The anatomical scan was followed either by a working memory task (total of 14.7 min) or an attention task (total of 16 min). Three subjects always performed the working memory task, and two subjects always performed the attention task. Over the next hour or so, subjects performed 4 runs of a sensorimotor task (described below, 4.25 min per run, total of 17 min), 2 runs of a breath-hold task (4.25 min) and 2 resting-state scans (fixation on a crosshairs, 4.25 min). These 8 scans were performed in a counterbalanced order. The entire scan session was repeated the following day. In the

Table 1
Description of hardware and sequences of the nine sites (10 scanners) participating in this study, five 1.5 T scanners, four 3 T scanners, and one 4 T scanner

Center	Abbreviation	Field Strength	Manufacturer	RF coil type	Functional sequence
Brigham and Women's	BWHM	3.0 T	GE	GE TR Research Coil	EPI
Duke/UNC	D40T	4.0 T	GE Nvi LX	TR quadrature head	Spiral
Duke/UNC	D15T	1.5 T	GE Nvi LX	TR quadrature head	Spiral
University of Iowa	IOWA	1.5 T	GE Signa CV/i	TR quadrature head	EPI
Mass. General Hospital	MAGH	3.0 T	Siemens Symphony Trio	TR quadrature head	EPI-Dual Echo
University of Minnesota	MINN	3.0 T	Siemens Symphony Trio	TR quadrature head	EPI
University of New Mexico	NMEX	1.5 T	Siemens Sonata	RO quadrature head	EPI
Stanford University	STAN	3.0 T	GE CV/NVi	Elliptical quadrature head	Spiral in/out
University of California, Irvine	UCIR	1.5 T	Philips/Picker	RO quadrature head	EPI
University of California, San Diego	UCSD	1.5 T	Siemens Symphony	TR quadrature head	EPI

present report, only data from the four sensorimotor runs and the two resting-state runs will be presented.

The functional data were collected using echo-planar (EPI) trajectories (7 scanners) or spiral trajectories (3 scanners) (see Table 1) (orientation: parallel to the AC–PC line, number of slices = 35, slice thickness = 4 mm, no gap, TR = 3000 ms, TE = 30 ms on the 3 T and 4 T scanners, 40 ms on the 1.5 T scanners, FOV = 22 cm, matrix = 64 × 64, voxel dimensions = 3.4375 mm × 3.4375 mm × 4 mm). MAGH (usually abbreviated MGH, see Table 1 for scanner abbreviations) employed a double-echo EPI sequence, and STAN employed a spiral in/spiral out sequence. All the spiral acquisitions were collected on General Electric (GE) scanners. The sensorimotor task produced 4 runs of 85 volumes each (85 TRs). The RF coils used varied with each scanner (Table 1).

Sensorimotor task

The sensorimotor (SM) task was designed by one of the authors (GHG) initially for other calibration purposes and employed a block design, with each block taking 10 TRs (30 s) beginning with 5 TRs (15 s) of rest (subject instructed to stare at fixation cross) and 5 TRs (15 s) of sensorimotor activity (see below). There were 8 full cycles of this followed by a 5 TR rest period at the end for a total of 85 TRs (4.25 min). During the active phase, subjects were instructed to tap their fingers bilaterally in synchrony with binaural tones, while watching an alternating contrast checkerboard. The checkerboard flash and tone presentation were simultaneous. The subjects were instructed to tap their fingers in an alternating finger-tapping pattern (index, middle, ring, little, little, ring, middle, index, index…) in synchrony with the tones and checkerboard flashes. The thumb was not used in this study. Each tone was 166 ms long with 167 ms of silence. The tone sequence utilized a dissonant series generated by a synthesizer (Midi notes 60, 64, 68, 72, 76, 80, 84, 88, 86, 82, 78, 74, 70, 66, 62, 58). The subjects' responses were recorded and monitored with the PST Serial Response Box (Psychology Software Tools, Inc., Pittsburgh, PA) (except for NMEX, which used a more ergonomically designed, custom built device).

Sensorimotor data analysis

The first step of image processing was accomplished using Analysis of Functional NeuroImage (AFNI) software (Cox, 1996). All large spikes in the data were removed from each sensorimotor run, and each run was motion corrected (i.e., spatially registered to the first volume of the run). The data were then slice-time corrected. A mean functional (T2*) image and also a detrended (2nd order polynomial) time series was created. The mean T2* image for each run was spatially normalized to an EPI canonical image in MNI space using tools available in SPM5b (http://www.fil.ion.ucl.ac.uk/spm/). This included affine transformations and 3 non-linear iterations. The spatial transformations were applied to the detrended time series data as well, and the time series was resampled at a 4 × 4 × 4 mm voxel size.

For level 1 statistical image analysis, Keith Worsley's package, FMRISTAT (Worsley et al., 2002; Liao et al., 2002) (http://www.math.mcgill.ca/keith/fmristat/) was employed. The first 2 volumes were discarded to allow for T1-saturation effects to stabilize. The statistical analysis was based on a linear model with correlated errors. The design matrix of the linear model, consisting of the off and on periods of the blocks, was first convolved with a hemodynamic response function modeled as a single gamma function [time to peak 4.7 s, FWHM = 3.8 s, Cohen (1997)] to produce the main regressor. In addition, the first derivative of this regressor was also included in the model to allow for some adjustment of the timing and shape of the response (Friston et al., 1998). Temporal drift was removed by adding a linear and quadratic contrast to the model. The correlation structure was modeled as a 1st degree autoregressive process. At each voxel, the autocorrelation parameters were estimated from the least squares residuals using the Yule–Walker equations, after a bias correction for correlations induced by the linear model. The autocorrelation parameters were first regularized by spatial smoothing, then used to 'whiten' the data and the design matrix. The linear model was then re-estimated using least squares on the whitened data. The parameter passed to level 2 statistical analysis was a common measure of effect size, i.e., the percent of variance accounted for (PVAF) defined as 100 * (sum of squares due to the main and derivative regressors)/(total sum of squares). The numerator is proportional to the variance due to the model, and the denominator is proportional to the total temporal variance about the mean. It is identical to $100 * r^2$, where r is the multiple Pearson correlation coefficient.

For a general discussion of the family of effect size metrics and the uses they support, see Rosenthal (1994) and Cohen (1998). If we had a single regressor, other effect sizes forms would have been appropriate, e.g., r or Cohen's D. Given that we have 2 regressors, and therefore 2 degrees of freedom in the numerator, multiple r^2 is an appropriate measure. We chose to multiply r^2 by 100 to express the result as percent of variance rather than proportion of variance. Model r^2 is closely related to an F value for the variance accounted for by both regressors:

$$F = [r^2 * (n - k - 1)]/[(1 - r^2) * k];$$

where n is the number of volumes ($n = 83$) and k is the number of regressors ($k = 2$). In this context, an F value can be thought of as a measure of contrast-to-fluctuation-noise-ratio (CFNR).

Slight modifications to FMRISTAT were required to produce PVAF, which will be referred to also as "activation effect size". For fMRI results, PVAF was given the sign of the beta weight for the main (i.e., non-derivative) regressor, although negative PVAF measures were generally not observed in this study.

Creation of ROIs

The SM task was expected to activate sensorimotor areas related to finger-tapping, auditory areas related to hearing tones and visual areas associated with observing the flashing checkerboard. We developed a set of "functionally defined" ROIs (Fig. 1) based on an analysis of the data in the present study. First, we performed a series of 1-sample t tests for each scanner and for each subject. The ROIs include only voxels that were activated at all 10 scanners with an uncorrected p value < 0.00001 and in all 5 subjects with an uncorrected p value < 0.00001. There were 8 ROIs identified: left and right motor cortex (LM and RM), left and right auditory cortex (LA and RA), left and right cerebellar cortex (LC and RC), bilateral visual cortex (BV) and the supplementary motor area (SM) (Fig. 1). For level II analysis, we brought forward the median PVAF value within each region for each run.

As discussed in Friedman et al. (2006), in a effort to reduce the number of variables to be analyzed at this stage, PVAF data from all 8 ROIs were entered into a factor analysis (SPSS, Inc). Only 1

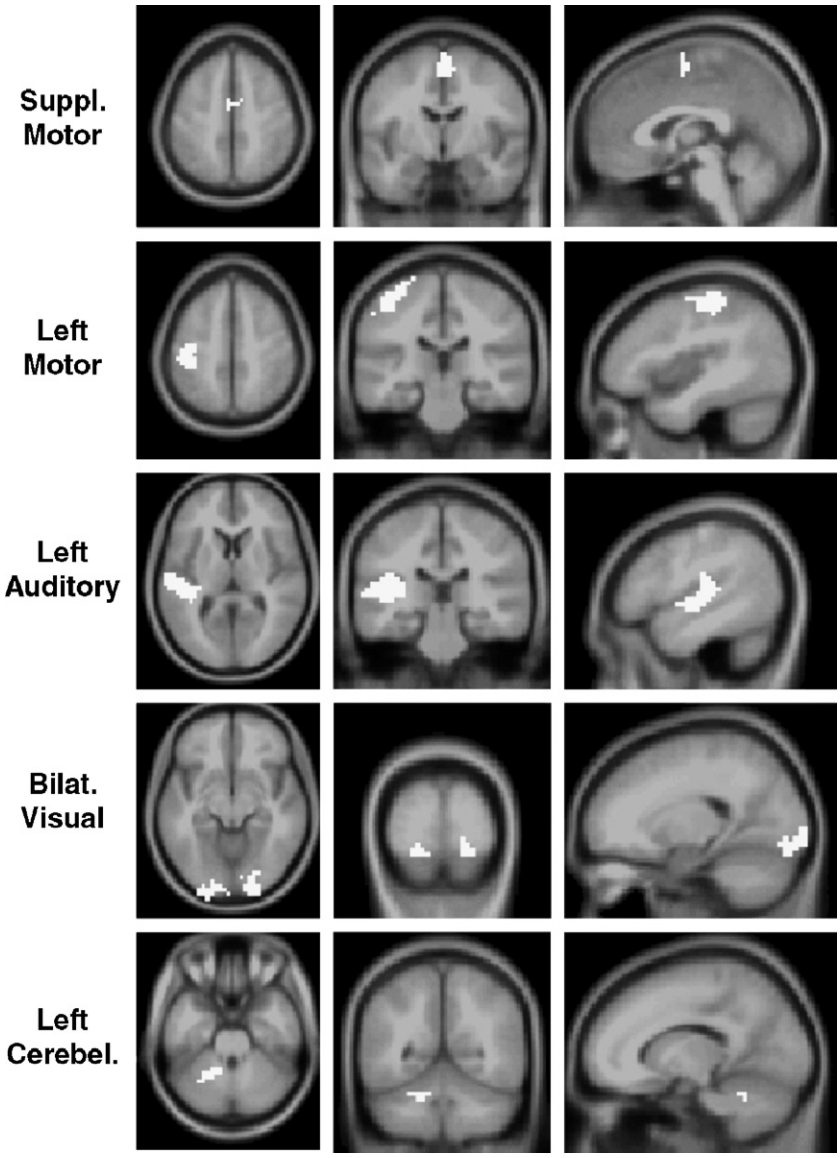

Fig. 1. Five of 8 ROIs employed in this study. ROIs are shown in white. For each ROI, axial, coronal and sagittal views are presented. The remaining 3 ROIs (right motor, right auditory and right cerebellum) were comparable to the contralateral ROIs shown in this figure.

factor (eigenvalue >1.0) was found, and all 8 ROIs were substantially weighted on this factor. However, the initial communality for the bilateral visual cortex ROI was substantially lower (0.53) than that for the other 7 ROIs (mean=0.81, SD=0.05; minimum=0.71). This may have been due to the fact that, in this initial study, display luminance and visual angle was somewhat dissimilar across scanners (data not shown). For these reasons, data from the visual cortex region were dropped from further analyses. For overall assessment of factors explaining variance in PVAF, the PVAF values for the remaining 7 ROIs were averaged. However, data on each of these 7 ROIs are also presented below.

Measuring signal-to-fluctuation-noise-ratio (SFNR)

We compared four measures of SFNR on effectiveness in reducing scanner-to-scanner variability of activation. For all four measures, SFNR was computed based on a time series of (fully preprocessed) T2* volumes. The four types varied based on the type of tissue included in the estimate (gray matter vs. white matter), whether the SFNR estimate was calculated from data from the resting-state task or the sensorimotor task and on the method of calculation. The gray matter (GM) and white matter (WM) masks were taken from the "discrete model" of the "Anatomical Model of Normal Brain" described at http://www.bic.mni.mcgill.ca/brainweb/anatomic_normal.html.

The four types of SFNR are described below. For SFNR Types 1 and 2, "signal" refers to the mean image intensity in a region over time (multiple T2* image volumes). "Fluctuation noise" was defined as the standard deviation over time after a 2nd order polynomial detrending of the same time series.

SFNR-Type-1-GM-R

For SFNR Type 1, all gray matter (GM) voxels were included. The SFNR was calculated separately for each of the

two resting-state scans. The two SFNR images were averaged to create a single SFNR image representing both resting-state runs. The median SFNR value in gray matter from this image was taken as a scalar value representing SFNR for this entire visit.

SFNR-Type-2-WM-R

This type of SFNR estimate was calculated exactly as SFNR Type 1 estimates except that a WM mask was employed rather than a GM mask.

For SFNR Types 3 and 4, the signal was taken as the estimate of the intercept from the level 1 analyses provided by FMRISTAT in the context of fitting all the relevant regressors and whitening the noise and model. In practice, this distinction is not very important because the mean of the data and the intercept of the model are almost identical. More critical is that the noise estimate is taken as the standard deviation of the model residuals also provided by FMRISTAT.

SFNR-Type-3-GM-SM-Resid

In this case, a GM mask was employed when calculating SFNR for each of the 4 runs of each visit.

SFNR-Type-4-WM-SM-Resid

This type of SFNR estimate was calculated exactly as Type 3 estimates except that a WM mask was employed rather than a GM mask.

Statistical analysis

Data were available for 5 subjects at 10 scanners. Each subject was scanned on 2 visits, and there were 4 SM and 2 resting-state runs per visit. Run effects were negligible and were consequently unmodeled (considered as part of the residual variance). Almost all analyses were performed in SAS (Cary, NC) and most employed the PROC MIXED procedure (Littell et al., 1996), a very sophisticated and flexible program for fitting mixed models. Restricted Maximum Likelihood Estimation (ReML) was used to estimate variance components. Degrees of freedom for fixed effects were estimated using the calculations detailed by Kenward and Roger (1997). The scanners were divided into low-field strength scanners (five scanners at 1.5 T) and high-field strength scanners (four 3 T and one 4 T scanner).

Reducing scanner-to-scanner variance by controlling for SFNR

To assess the effect of scanner (within field) without correction, a mixed model ANOVA was employed. PVAF was the dependent variable. Scanner was modeled as a fixed effect. Subject, visit, subject-by-visit, site-by-subject, site-by-visit and site-by-subject-by-visit were modeled as random effects. To assess the effectiveness of various SFNR estimates, each estimate was entered singly as a covariate into the above model. Thus, there were five types of mixed model analyses conducted, 1 ANOVA and 4 ANCOVAs. These five analyses were conducted separately for low-field scanners and high-field scanners, and for 8 ROIs (LA, RA, LC, RC, LM, RM, SM and the average, AV). Thus, there were 80 mixed model ANOVAs or ANCOVAs conducted.

It is important to note the analysis structure here. For each visit, each subject had only a single SFNR-Type-1-GM-R estimate and a single SFNR-Type-2-WM-R estimate. So all four runs of the SM task were corrected with the same resting-state SFNR covariate. However, for SFNR Types 3 and 4, each individual run of the SM task was corrected by an SFNR estimate from that same run.

Criteria for ranking the four SFNR types

If the goal is to reduce scanner effects, then obviously statistically significant F tests of the scanner effect are bad, and any step which reduces such an effect to non-statistically significant levels ($p > 0.05$) is good. Therefore, the number of statistically significant F tests for the scanner effect in each model was calculated. Furthermore, the less variable the estimated scanner means were for each model, the better the method is at reducing variance due to scanner. Therefore, the coefficient of variation (CV) was calculated for the estimated scanner means for each model.

The relationship between SFNR and activation effect size (PVAF) was assessed with a random-coefficients model (Littell et al., 1996). The random-coefficients model is a form of hierarchical linear modeling. At the first level, slopes and intercepts for the relationships in question are estimated for each subject. The first level variables are then analyzed as random variables and a mean (across subjects) slope and intercept are estimated. Once the mean slope and mean intercept are in hand, it is straightforward to predict PVAF at any arbitrary level of SFNR (see below).

Results

To illustrate the analytic approach employed, we present a detailed analysis of one of the most effective SFNR estimates (SFNR-Type-1-GM-R). After this detailed presentation, for this single SFNR estimate, comparative data on the various SFNR estimates are presented.

SFNR-Type-1-GM-R across scanners

SFNR-Type-1-GM-R levels differed markedly across scanners, even within a field (Fig. 2). The F value for the scanner effect on SFNR Type 1 for low-field scanners was 103.2 ($df = 4$, 16.1, $p < 0.0001$), and the F value for the high-field scanners was 15.0 ($df = 4$, 16.9, $p < 0.0001$).

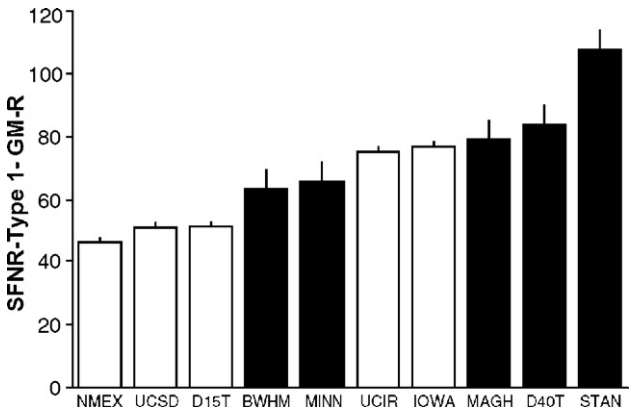

Fig. 2. SFNR-Type-1-GM-R across scanners. The open bars indicate low-field scanners and the dark-filled bars indicate high-field scanners. The error bars represent the standard error. Both the means and the standard errors of these means were estimated using the LSMEANs option of SAS PROC MIXED.

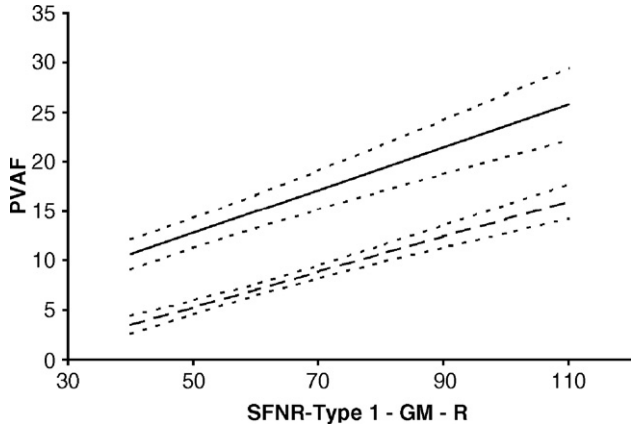

Fig. 3. A: Plot of the relationship between SFNR-Type-1-GM-R (abscissa) and PVAF (ordinate, activation effect size) across all 5 subjects, as estimated by the random-coefficients regression of PVAF onto SFNR described in the text. The dashed line represents the relationship for low-field scanners and the solid line represents the relationship for high-field scanners. The dotted lines are standard errors for these regression lines.

Relationship between SFNR-Type-1-GM-R and activation effect size

The relationship between SFNR-Type-1-GM-R and activation effect size (PVAF) is illustrated in Fig. 3. This relationship was statistically significant for both low-field and high-field scanners (low field: $F=28.8$, $df=1,4$, $p=0.006$; high field: $F=17.6$, $df=1,4$, $p=0.014$). The slopes of the relationship for each subject are presented in Table 2. The mean slope of the relationship for low-field scanners (0.18) was quite similar to the slope for high-field scanners (0.21). Thus, for every unit of increase in SFNR, activation effect size (PVAF) is increased by approximately 0.20 units.

Scanner effects before and after adjustment with SFNR-Type-1-GM-R

The mean SFNR-Type-1-GM-R for low-field scanners was 60, and the ANCOVA compared the data after each site was adjusted as if it had been all collected at a single site with an SFNR-Type-1-GM-R of 60. For high-field scanners, SFNR-Type-1-GM-R was 79.9 [very close to the value at MAGH (79.0)]. *F* values for the scanner effect before and after adjustment with SFNR-Type-1-GM-R are presented in Figs. 4 (low field) and 5 (high field). For low-field scanners, in the unadjusted case, the *F* values were statistically significant for 8 of 8 measures. After adjustment with SFNR-Type-1-GM-R, none of the *F* values was statistically significant. For high-field scanners, in the unadjusted case, the *F* values were statistically significant for 6 of 8 measures. After adjustment with SFNR-Type-1-GM-R, none of the *F* values was statistically significant.

Table 2
Slope estimates for each subject relating SFNR to PVAF from low-field and high-field scanners

Field strength	Subject	Slope estimate	*df*	*t* value	*p* value
Low	1	0.13	38	4.6	<0.0001
	2	0.22	38	5.5	<0.0001
	3	0.09	38	2.8	0.009
	4	0.17	38	6.3	<0.0001
	5	0.28	38	7.0	<0.0001
Mean slope		0.18			
High	1	0.08	38	1.5	0.2
	2	0.39	38	7.8	<0.0001
	3	0.22	38	2.4	0.02
	4	0.14	38	2.1	0.04
	5	0.23	34	5.4	<0.0001
Mean slope		0.21			

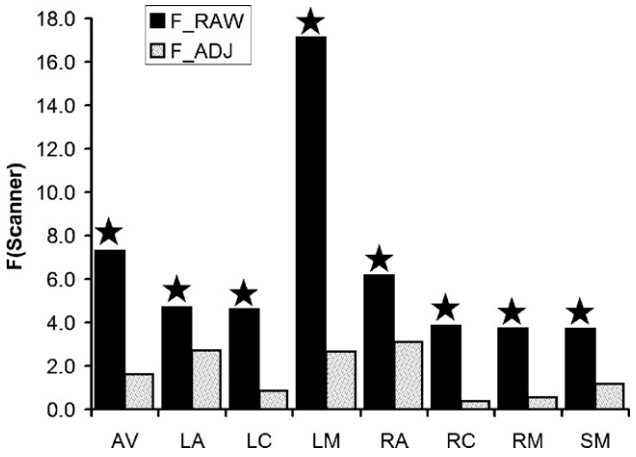

Fig. 4. Plot of the *F* values for the scanner effect within the low-field-scanners over the 8 measures (average and 7 ROIs). The black filled bars represent data before covariate adjustment with SFNR-Type-1-GM-R (F_Raw) and gray spotted bars represent data after covariate adjustment with SFNR-Type-1-GM-R (F_Adj). Stars indicate statistical significance ($p<0.05$).

Estimated scanner means before and after adjustment with SFNR-Type-1-GM-R

Figs. 6 and 7 show the least squares estimated means for each site before and after adjustment for SFNR-Type-1-GM-R, for low-field and high-field scanners, respectively. For low-field scanners, the coefficient of variation among scanner mean estimates before adjustment was 0.40. After adjustment for SFNR Type 1, the coefficient of variation was 0.22, a reduction in variation of 44%.

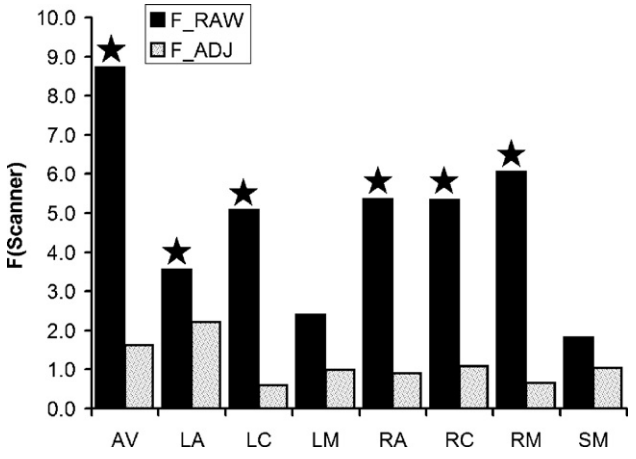

Fig. 5. Plot of the *F* values for the scanner effect within the high-field scanners over the 8 measures (average and 7 ROIs). For remaining details, see Fig. 4.

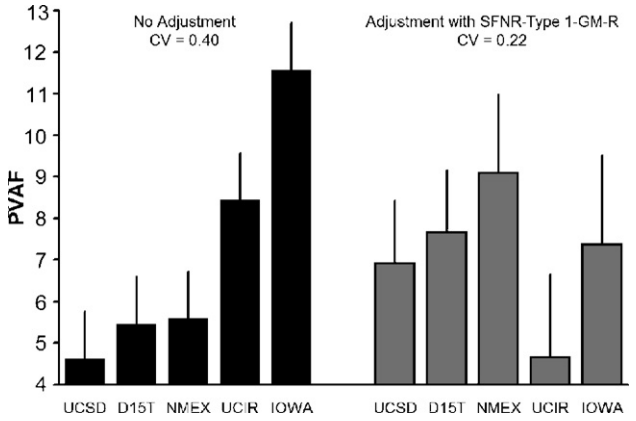

Fig. 6. Estimated scanner mean PVAF results for low-field scanners before and after adjustment with SFNR-Type-1-GM-R. The black bars represent data before covariate adjustment and the gray bars represent data after covariate adjustment with SFNR-Type-1-GM-R. The error bars are standard errors.

For high-field scanners, the coefficient of variation among scanner mean estimates before adjustment was 0.26. After adjustment for SFNR Type 1, the coefficient of variation was 0.19, a reduction in variation of 27%.

For low-field scanners (Fig. 6), the 3 "low-PVAF sites" (USCD, D15T and NMEX) all show increases in estimated PVAF after correction, whereas the two "high-PVAF sites" (UCIR and IOWA) show decreases in estimated PVAF after correction, thus leading to a decrease in the CV after correction. However, after covariate adjustment, UCIR appeared to be somewhat overcorrected. The cause for this is explored below. For high-field scanners (Fig. 7), the 2 "low-PVAF sites" (BWHM and MINN) show increases in estimated PVAF after correction, whereas the two "high-PVAF sites" (D40T and STAN) show decreases in estimated PVAF after correction, thus leading to a decrease in the CV after correction.

Is UCIR an outlier site?

As noted above, the UCIR site appears to be somewhat over-corrected after adjustment for variation in SFNR-Type-1-GM-R. Fig. 8 plots the relationship between SFNR-Type-1-GM-R and PVAF in one strongly activated region, the right auditory cortex

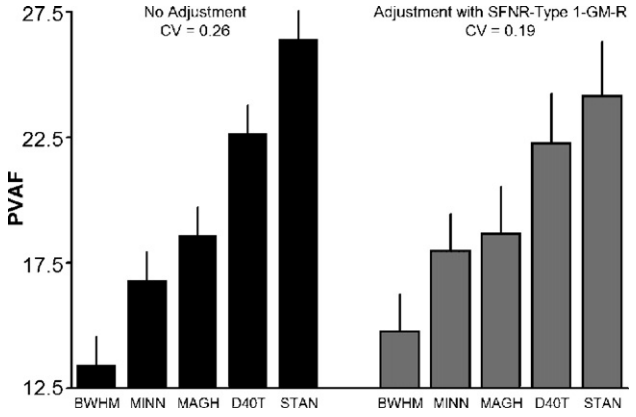

Fig. 7. Estimated scanner mean PVAF results for high-field scanners before and after adjustment with SFNR-Type-1-GM-R. See Fig. 6 for details.

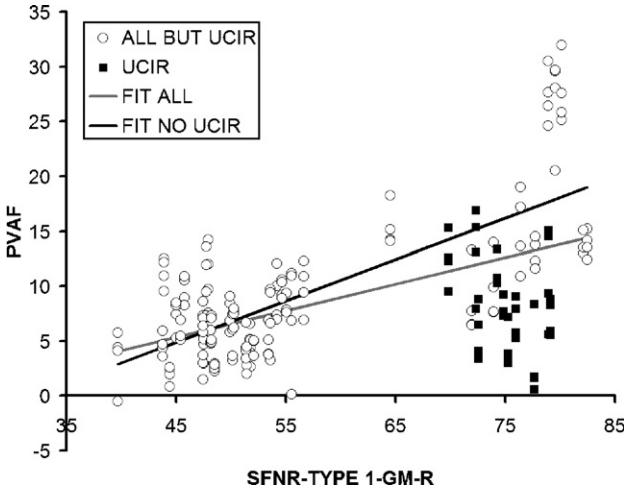

Fig. 8. Scatterplot relating SFNR-Type-1-GM-R to PVAF values for each run for the RA ROI. Since there were 5 subjects studied on two visits with 4 runs per visit, there are 40 points per site. The black squares represent the data from the UCIR site. All other data are indicated by a open circle. The gray line (smaller slope) is the linear fit to all the data and the black line is the linear fit to the data with UCIR excluded.

("RA"). Note that although UCIR (black squares) data generally have a high SFNR Type 1, PVAF in right auditory cortex is not proportionately increased. Note that the regression accounts for much more variance when UCIR is excluded (black line, 54%) than when UCIR is included (gray line, 29%). Examination of similar scatterplots from other ROIs indicates that this pattern only holds for the LA and RA regions suggesting a weak response to auditory stimulation at UCIR. UCIR employed the standard FBIRN equipment and procedures for the auditory stimulation aspect of the task so there is no obvious explanation for this difference. One possibility is that some other quality of the environment, for example, some spectral property of the scanner noise, may have interfered with subjects perception of the tones presented. It would have been of interest to explore this issue further, but this scanner has since be decommissioned and replaced.

The following questions are raised by this UCIR data: is this site an outlier? Should it be included when choosing among SFNR types? For the FBIRN project going forward, this scanner has been replaced so the question is somewhat moot. Below, results are presented for low-field scanners with and without UCIR.

Are measures of SFNR statistically significant regressors for PVAF?

The F values and associated p values for the 4 measures of SFNR for low-field scanners (with and without UCIR) and high-field scanners are presented in Table 3. Ten of twelve tests were

Table 3
F values for the regression of SFNR Types on PVAF for low-field scanners, low-field scanners without UCIR (No UCIR) and for high-field scanners

Type	Label	Low		No UCIR		High	
		F	p	F	p	F	p
1	GM-R	5.3	0.03	5.0	0.03	0.8	0.4
2	WM-R	4.5	0.04	4.6	0.04	1.3	0.3
3	GM-SM-Resid	61.6	<0.0001	43.4	<0.0001	81.2	<0.0001
4	WM-SM-Resid	60.9	<0.0001	42.9	<0.0001	73.6	<0.0001

Table 4
Mean SFNR estimates for low-field and high-field scanners

Field	SFNR-Type-1-GM-R	SFNR-Type-2-WM-R	SFNR-Type-3-GM-SM-Resid	SFNR-Type-4-WM-SM-Resid
Low	60.0	62.0	101.4	107.4
High	79.7	93.1	124.0	150.2

statistically significant. Thus, one can conclude that, generally speaking, SFNR is statistically significant as a covariate for PVAF when comparing scanner effects. The *F* values for the resting-state measures (GM-R and WM-R) are probably low because of the fewer degrees of freedom for these tests since each visit had a single SFNR-GM-R estimate but each run (4 per visit) had its own SFNR Types 3 and 4. We also tested for the possibility of a curvilinear relationship by adding a quadratic component to the models. As a general matter, there was almost no difference at all between the *F* value for the scanner effect with or without the quadratic component (data not shown), so it appears that a linear model will suffice.

Comparative analysis of 4 SFNR types—mean values

White matter estimates of SFNR were always greater than the corresponding gray matter estimates (Table 4). SFNR estimates based on data from the resting-state (Types 1 and 2) were lower than SFNR estimates based on the SM task (Types 3 and 4). Statistical analysis revealed that GM had a lower mean intensity during the rest task than during the SM task at 9 of 10 sites (all p values <0.01 for all sites but BWHM), and that at 10 of 10 sites GM had a higher standard deviation during rest than during the sensorimotor task. The intercept estimates for SFNR types 3 and 4 were essentially indistinguishable from the simple intensity means.

The mean values in Table 4 are what the adjusted PVAF estimates are corrected to in the ANCOVA models. For example, when high-field scanners are adjusted for SFNR-Type-3-GM-SM-Resid, the PVAF values are adjusted to a level that would be predicted at imaginary average scanner with an SFNR-Type 3-GM-SM-Resid of 124.

Comparative analysis of SFNR types

The values in Table 5 should now be familiar. The "% change" columns represent percent change in CV from the unadjusted case. A large positive value means that the SFNR Type in that row reduced scanner variation by a large amount. The "*N* significant" columns is the number of *F* tests for scanner that were statistically significant for each of the 7 ROIs plus the average of the ROIs. The highest possible value here is 8. The smaller the number, the more effective an SFNR Type is in reducing scanner effects across the ROIs.

How one rates the effectiveness of the various SFNR Types depends on several factors: is one interested primarily in low- or high-field scanners? Is one convinced or unconvinced that UCIR is an outlier? Is one interested in avoiding collecting separate resting-state data?

For all low-field scanners only, SFNR-Type-1-GM-R is the clear winner. Variation among the scanners (CV) is reduced by 44% (UCIR included) or 67% (UCIR excluded). None of the 8 *F* tests for scanner effects was statistically significant after adjusting for this SFNR Type. For high-field scanners only, SFNR-Type-3-GM-SM-Resid is the clear winner since adjustment for this SFNR Type reduced variation among scanners by approximately 53%. However, in terms of the number of *F* tests for scanner that were significant after adjustment, all SFNR types resulted in 0 statistically significant *F* values, so SFNR Types are equivalent for high-field scanners in this sense.

For the FBIRN Project going forward, only high-field scanners will be included. However, avoidance of the burden of collecting extra resting-state runs is an important goal. In this case, SFNR-Type-3-GM-SM-Resid looks like a very good choice. It leads to a reduction in CV of 53%, and it is computed without the burden of collecting an additional 8.5 min of data.

Effect of adjustment for SFNR-Type-3-GM-SM-Resid on variance due to subject and residual variance

To further explore the effects of adjustment for SFNR, we performed a variance components analysis to assess the effects of adjustment for our chosen SFNR-Type, i.e., SFNR-Type-3-GM-SM-Resid. We modeled PVAF as a function of 3 random effects: site, subject and visit. We compared the estimate of the variance due to these effects and the residual variance in a model without any covariates and in a model with SFNR-Type 3 as a covariate (Table 6). The variance due to scanner was reduced markedly (low field, No UCIR: 75%, high field: 81%). The variance due to subject was quite small in the low-field case to begin with (1.8), and covariate adjustment led to a minimal increase in this value to 2.4. In the high-field case, there was a substantial variance due to subject, and covariate adjustment reduced this variance by 40%. Residual variance was also decreased by 26% to 29% with covariate adjustment. Thus, not only does covariate adjustment decrease the site effects, but it also improves the models in other

Table 5
Comparison among SFNR types in reducing activation effect size differences due to scanners

Label	Low			Low − No UCIR			High		
	CV[a]	% Change[b]	*N* significant[c]	CV	% Change	*N* significant	CV	% Change	*N* significant
	0.40		8	0.47		7	0.26		6
SFNR-Type-1-GM-R	0.22	44	0	0.15	67	0	0.19	27	0
SFNR-Type-2-WM-R	0.31	23	0	0.29	38	0	0.16	38	0
SFNR-Type-3-GM-SM-Resid	0.32	20	5	0.19	59	0	0.12	53	0
SFNR-Type-4-WM-SM-Resid	0.43	−6	5	0.36	24	2	0.17	35	0

[a] "CV" is coefficient of variation.
[b] "% change" is the percent reduction in the CV among site means when an SFNR type is used.
[c] "*N* significant" is the number of measures (7 ROIs and the average of the 7 ROIs) for which there was a statistically significant *F* test.

Table 6
Variance components analysis based on adjustment with SFNR-Type-3-GM-SM-Resid

Covariance parameter	No adjustment	With adjustment	% decrease
Low−No UCIR[a]			
Scanner	10.0[b]	2.5	74.6
Subject	1.8	2.4	−30.6
Visit	0.0	0.1	
Residual	8.5	6.0	28.9
High			
Scanner	24.3	4.7	80.8
Subject	32.9	19.8	39.8
Visit	4.1	2.7	34.7
Residual	42.3	31.4	25.9

[a] "Low−No UCIR" refers to the set of low-field scanners minus UCIR.
[b] Values are actual raw variance estimates.

ways, including a reduction of variance due to subjects in the high-field case and an overall reduction in residual (i.e., unexplained) variance.

Using ROI-specific estimates of SFNR

The SFNR estimates employed above for GM were based on all GM voxels throughout the brain. It was of interest to determine if more spatially specific estimates from each ROI would improve SFNR adjustment for site effects. For this analysis, we employed on SFNR-Type-3-GM-SM-Resid. The results suggest that the switch from spatially general estimates of SFNR to spatially specific estimates did not make a consistent positive or negative difference. Among the 14 tests conducted (7 ROIs at 2 field strengths), ROI-based estimates of SFNR resulted in a lower F (scanner) than a GM-based estimate in 8 cases (4 at low field and 4 at high field). Possibly, the added spatial specificity of the ROI-based estimates was counteracted by the decreased stability of these estimates: GM estimates were based on more than 14,000 voxels whereas the ROI-based estimates were based on anywhere from 18 to 176 voxels.

Discussion

The strategy of reducing scanner effects in multicenter studies by using measures of SFNR as covariates in ANCOVA designs is clearly successful. SFNR estimates are substantially different between sites within (and between) field strength. SFNR estimates are generally correlated with our chosen dependent variable, activation effect size (PVAF). Covarying for any one of several SFNR types reduced F tests for scanner effects from highly statistically significant to statistically non-significant. Variations among estimated scanner means were also markedly reduced.

Why does SFNR differ between sites even within a field strength? Among the low-field scanners, clearly UCIR and IOWA have the highest SFNR. Three of the four sites with the highest SFNR all used apodization (k-space) filters; the filter at the IOWA site was particularly severe (Friedman et al., 2006). Such filters impart additional smoothness to the data and would be expected to increase SFNR (Friedman and Glover, 2006; Friedman et al., 2006). Many other factors can obviously affect SFNR, for example the shape of k-space, the readout time, the design and tuning of the head coil, etc. Among the high-field group, the STAN site had a markedly elevated SFNR compared to the other scanners, including a 4 T. This site employed a spiral-in-out "double-echo" acquisition which would be expected to produce an elevated SFNR. Furthermore, the STAN site used a custom head coil with very high sensitivity. The next highest SFNR estimate was for a 4 T scanner (D40T). MAGH used a double-echo EPI sequence which would also provide elevated SFNR.

SFNR was linearly related to activation effect size, as evidenced by the statistically significant F tests for the covariates (Table 3). A decrease in noise, all other things being equal, will always lead to an increase in SFNR, contrast-to-fluctuation-noise-ratio (CFNR) and the value of a t statistic or effect size measure as the noise is in the denominator of each of these measures. A theoretical and empirical basis for a relationship between SFNR and t tests for activation has been presented by Parrish et al. (2000). There was no evidence that the relationship between SFNR and PVAF was curvilinear. Given that SFNR will pick up unwanted variance due to scanner and is related to the dependent variable chosen here (activation effect size), it makes an excellent covariate for the purpose of reducing scanner effects.

The UCIR site had a somewhat unique SFNR/activation effect size relationship for auditory cortex (left and right). This scanner was a somewhat older model and was the only Picker scanner in the study. Although this site had a high SFNR, the activation effect sizes were quite low for the auditory cortex activations. Such a result could occur if the auditory stimulation at this site was relatively weak, but we have been assured by the investigators at this site that the same equipment was used as other sites as well as the same protocol employed for volume adjustment. As suggested above, maybe some other quality of the environment, for example, some spectral property of the scanner noise, interfered with subjects perception of the tones presented. Gaab et al. (in press) have recently reported a marked effect of scanner background noise on auditory activation in the superior temporal gyrus. Since this scanner has been decommissioned, further examination of this issue is impossible. It is the view of the present authors that it is reasonable to consider the UCIR site as an outlier for the purposes of the analysis presented here.

The results suggest that SFNR-Type-3-GM-SM-Resid is the best SFNR estimate to use in the present context. It does not require the collection of an additional resting-state scan, which is a major advantage considering the multiple competing interests in multicenter studies for scanner time. It is highly effective at reducing scanner effects to a statistically non-significant level and at reducing variation among estimated scanner means. Since it is based on the model residuals after removal of variance due to the activation task, it will not be affected by tasks with widespread activation. In addition, it is quite effective at reducing subject-related variance in the high-field case and reducing the amount of unexplained variance for both low-field and high-field scanners. These latter effects (reducing in subject-related and unexplained variance) may be a reason to consider including SFNR estimates even in unicenter studies.

As a general matter, SFNR estimates based on gray matter were more effective than SFNR estimates based on white matter. To understand this difference, it is important to take a closer look at noise in fMRI time series. Noise in fMRI time series is generated by 2 types of sources: non-physiological sources and physiological sources (Edelstein et al., 1986; Kruger and Glover, 2001; Kruger et al., 2001; Triantafyllou et al., 2005). At 1.5 T and above, non-

physiological noise comes primarily from thermal "Johnson" noise from the subject and scanner noise related to instabilities in the scanner hardware, such as minor fluctuations in the amplitude of an RF pulse over time. Physiological noise consists of a BOLD-related component, picking up background neuronal and metabolic activity, and a non-BOLD-related component from respiratory, cardiac and brain pulsation artifacts. At 3 T, non-physiological noise is similar in GM and WM (Kruger and Glover, 2001). In contrast, physiological noise is on average about 2 times higher in gray matter than in white matter regions. Furthermore, physiological noise is consistently larger than non-physiological noise in GM at 3 T. At 3 T, BOLD-related physiological noise is approximately 1.9× larger in GM than in WM and is approximately 2.0× larger than the non-BOLD-related physiological noise (Kruger and Glover, 2001). Therefore, at 3 T at least, temporal noise in GM is likely to be quite highly weighted toward BOLD-related physiological noise. This sensitivity to this type of noise in GM may be related to the increased effectiveness of SFNR estimates from GM in reducing scanner and subject effects at 3 T. Our results in Table 5 are in accord with these characteristics.

Post hoc adjustment for SFNR is not the only method to reduce scanner effects on activation. Obviously, steps which minimize SFNR variation among sites prior to study initiation will reduce the need for this kind of correction. The presence of various apodization (*k*-space) filters, which impart image smoothness, contributes to site differences in SFNR. The fact that some sites in the present study employed double-echo acquisitions also creates scanner differences in SFNR. Finally, different head coils at different sites can also increase site differences in SFNR. Despite these obvious site differences, the proposed methods utilizing SFNR successfully eliminated significant site effects.

Although it would be ideal to be able to merge data from 1.5 T and 3 T scanners, we are somewhat reluctant to recommend this now. Two classic papers, Gati et al. (1997) and Ugurbil et al. (1999), have shown that cerebral elements (capillaries, venules, veins) contribute differentially to T2* images collected at 1.5 T and at 3 T and higher. Therefore, at a fundamental level, the data from these two field strengths are qualitatively different. In a recent paper just submitted, Dr. Friedman has provided data on the size of the signal (percent signal change) from 1.5 T scanners and 3 T scanners included in the present study. 3 T scanners have markedly higher signal leading to a step-like change in contrast-to-noise ratios. The constituents of the noise also differ, with greater physiological noise at high-field (Kruger and Glover, 2001; Kruger et al., 2001). Marked differences in susceptibility artifact and geometric distortion with EPI at the two field strengths also present difficulties in merging across field strength. High-field scanners have a reduced T2-relaxation time, which means that the MR signal is decaying more rapidly, leading to a wider pixel point spread function at high field (Farzaneh et al., 1990). Consequently, until methods are developed to adequately address these differences, we are reluctant to recommend that data from low and high-field scanners be merged at this point.

In the present study, we employed activation effect size (PVAF) as our measure of analysis from level 1 statistical analysis. This parameter carries with it influence from both signal and noise in the data. By homogenizing this parameter across sites, we are homogenizing the CNR and statistical significance of activations across sites. Another advantage of this measure is that it can be used as a measure of effect size, allowing for direct comparison between any type of fMRI contrast from any tasks, and it can be used to conduct power analyses to determine the amount of time one needs to devote to a particular task–contrast combination.

Acknowledgments

This research was supported by a grant [#1 U24 RR021992] to the Functional Imaging Biomedical Informatics Research Network (FBIRN, http://www.nbirn.net), that is funded by the National Center for Research Resources (NCRR) at the National Institutes of Health (NIH).

The members of the FBIRN project all deserve acknowledgement for their significant efforts, but unfortunately, they are too numerous to mention. Please visit http://www.nbirn.net/TestBeds/Function/index.htm for more information regarding key personnel. We would identify two people for special acknowledgement: Dr. Jessica Turner, the project coordinator, worked tirelessly to facilitate the entire project; and the FIRST BIRN principal investigator, Dr. Steven Potkin, who provided vision and leadership throughout.

References

Bevan, J.S., Atkin, S.L., Atkinson, A.B., Bouloux, P.M., Hanna, F., Harris, P.E., James, R.A., McConnell, M., Roberts, G.A., Scanlon, M.F., et al., 2002. Primary medical therapy for acromegaly: an open, prospective, multicenter study of the effects of subcutaneous and intramuscular slow-release octreotide on growth hormone, insulin-like growth factor-I, and tumor size. J. Clin. Endocrinol. Metab. 87 (10), 4554–4563.

Brada, M., Hoang-Xuan, K., Rampling, R., Dietrich, P.Y., Dirix, L.Y., Macdonald, D., Heimans, J.J., Zonnenberg, B.A., Bravo-Marques, J.M., Henriksson, R., et al., 2001. Multicenter phase II trial of temozolomide in patients with glioblastoma multiforme at first relapse. Ann. Oncol. 12 (2), 259–266.

Casato, M., Saadoun, D., Marchetti, A., Limal, N., Picq, C., Pantano, P., Galanaud, D., Cianci, R., Duhaut, P., Piette, J.C., et al., 2005. Central nervous system involvement in hepatitis C virus cryoglobulinemia vasculitis: a multicenter case–control study using magnetic resonance imaging and neuropsychological tests. J. Rheumatol. 32 (3), 484–488.

Casey, B.J., Cohen, J.D., O'Craven, K., Davidson, R.J., Irwin, W., Nelson, C.A., Noll, D.C., Hu, X., Lowe, M.J., Rosen, B.R., et al., 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. NeuroImage 8 (3), 249–261.

Chang, L., Lee, P.L., Yiannoutsos, C.T., Ernst, T., Marra, C.M., Richards, T., Kolson, D., Schifitto, G., Jarvik, J.G., Miller, E.N., et al., 2004. A multicenter in vivo proton-MRS study of HIV-associated dementia and its relationship to age. NeuroImage 23 (4), 1336–1347.

Cohen, M.S., 1997. Parametric analysis of fMRI data using linear systems methods. NeuroImage 6 (2), 93–103.

Cohen, J., 1998. Statistical Power Analysis for the Behavioral Sciences.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.

Edelstein, W.A., Glover, G.H., Hardy, C.J., Redington, R.W., 1986. The intrinsic signal-to-noise ratio in NMR imaging. Magn. Reson. Med. 3 (4), 604–618.

Ewers, M., Teipel, S.J., Dietrich, O., Schonberg, S.O., Jessen, F., Heun, R., Scheltens, P., Pol, L.V., Freymann, N.R., Moeller, H.J., et al., 2006. Multicenter assessment of reliability of cranial MRI. Neurobiol. Aging 27 (8), 1051–1059.

Farzaneh, F., Riederer, S.J., Pelc, N.J., 1990. Analysis of T2 limitations and off-resonance effects on spatial resolution and artifacts in echo-planar imaging. Magn. Reson. Med. 14 (1), 123–139.

Freimer, N., Sabatti, C., 2003. The human phenome project. Nat. Genet. 34 (1), 15–21.

Friedman, L., Glover, G.H., 2006. Report on a multicenter fMRI quality assurance protocol. J. Magn. Reson. Imaging 23 (6), 827–839.

Friedman, L., Glover, G.H., Krenz, D., Magnotta, V., FIRST BIRN, 2006. Reducing Scanner-to-scanner variability of activation in a multi-center fMRI study: role of smoothness equalization. NeuroImage (electronic publication ahead of print, 2006 Jul 26).

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-related fMRI: characterizing differential responses. NeuroImage 7 (1), 30–40.

Gaab, N., Gabrieli, J.D.E., Glover, G.H., in press. Assessing the influence of scanner background noise on auditory processing—I. An fMRI study comparing three experimental designs with varying degrees of scanner noise. Hum. Brain Mapp.

Gati, J.S., Menon, R.S., Ugurbil, K., Rutt, B.K., 1997. Experimental determination of the BOLD field strength dependence in vessels and tissue. Magn. Reson. Med. 38 (2), 296–302.

Glover, G.H., Lai, S., 1998. Self-navigated spiral fMRI: interleaved versus single-shot. Magn. Reson. Med. 39 (3), 361–368.

Insel, T.R., Volkow, N.D., Landis, S.C., Li, T.K., Battey, J.F., Sieving, P., 2004. Limits to growth: why neuroscience needs large-scale science. Nat. Neurosci. 7 (5), 426–427.

Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53 (3), 983–997.

Kruger, G., Glover, G.H., 2001. Physiological noise in oxygenation-sensitive magnetic resonance imaging. Magn. Reson. Med. 46 (4), 631–637.

Kruger, G., Kastrup, A., Glover, G.H., 2001. Neuroimaging at 1.5 T and 3.0 T: Comparison of oxygenation-sensitive magnetic resonance imaging. Magn. Reson. Med. 45 (4), 595–604.

Liao, C.H., Worsley, K.J., Poline, J.B., Aston, J.A., Duncan, G.H., Evans, A.C., 2002. Estimating the delay of the fMRI response. NeuroImage 16 (3 Pt. 1), 593–606.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.S., 1996. SAS System for Mixed Models.

Lowe, M.J., Sorenson, J.A., 1997. Spatially filtering functional magnetic resonance imaging data. Magn. Reson. Med. 37 (5), 723–729.

O'Connor, P.W., Goodman, A., Willmer-Hulme, A.J., Libonati, M.A., Metz, L., Murray, R.S., Sheremata, W.A., Vollmer, T.L., Stone, L.A., 2004. Randomized multicenter trial of natalizumab in acute MS relapses: clinical and MRI effects. Neurology 62 (11), 2038–2043.

Parrish, T.B., Gitelman, D.R., LaBar, K.S., Mesulam, M.M., 2000. Impact of signal-to-noise on functional MRI. Magn. Reson. Med. 44 (6), 925–932.

Rosenthal., 1994. Parametric Measures of Effect Size.

Schnack, H.G., van Haren, N.E., Hulshoff Pol, H.E., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T., Huttunen, M., Murray, R., Kahn, R.S., 2004. Reliability of brain volumes from multicenter MRI acquisition: A calibration study. Hum. Brain Mapp. 22 (4), 312–320.

Silver, N.C., Barker, G.J., Miller, D.H., 1999. Standardization of magnetization transfer imaging for multicenter studies. Neurology 53 (5 Suppl. 3), S33–S39.

Stocker, T., Schneider, F., Klein, M., Habel, U., Kellermann, T., Zilles, K., Shah, N.J., 2005. Automated quality assurance routines for fMRI data applied to a multicenter study. Hum. Brain Mapp. 25 (2), 237–246.

Tabachnick, B.G., Fidell, L.S., 1989. Using Multivariate Statistics. 2nd Ed.

Thomason, M.E., Foland, L.C., Glover, G.H., 2006. Calibration of BOLD fMRI using breath-holding reduces group variance during a cognitive task. Hum. Brain Mapp (electronic publication ahead of print, 2006 Jul 26).

Triantafyllou, C., Hoge, R.D., Krueger, G., Wiggins, C.J., Potthast, A., Wiggins, G.C., Wald, L.L., 2005. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. NeuroImage 26 (1), 243–250.

Ugurbil, K., Ogawa, S., Kim, S.G., Chen, W., Zhu, X.H., 1999. Imaging brain activity using nuclear spins. In: Maraviglia, B. (Ed.), Magnetic Resonance and Brain Function: Approaches from Physics. IOS Press, Amsterdam, pp. 261–310.

Worsley, K.J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F., Evans, A.C., 2002. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.

Zou, K.H., Greve, D.N., Wang, M., Pieper, S.D., Warfield, S.K., White, N.S., Manandhar, S., Brown, G.G., Vangel, M.G., Kikinis, R., et al., 2005. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. Radiology 237 (3), 781–789.