◆ Human Brain Mapping 27:442–451(2006) ◆

Diagnosis of Single-Subject and Group fMRI Data with SPMd

Hui Zhang,[1] Wen-Lin Luo,[2] and Thomas E. Nichols[1]*

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan*
[2]*CBARDS, Merck & Co., Rahway, New Jersey*

Abstract: Except for purely nonparametric methods, statistical methods depend on assumptions about the distribution of the data studied. While these assumptions are easily checked for a single univariate dataset with diagnostic plots, in the massively univariate model used with functional MRI (fMRI) it is impractical to check with a massive number of plots. In previous work we have demonstrated how to diagnose model assumptions and lack-of-fit for single-subject fMRI models using a working assumption of independent errors; our work depended on images and time series of summary statistics that, when simultaneously viewed dynamically, identify problem scans and voxels. In this article we extend our previous work to account for temporal autocorrelation in single-subject models and show how analogous methods can be used on group models where multiple subjects are studied. We apply these methods to the single-subject Functional Image Analysis Contest (FIAC) data and find several anomalies, but none that appear to invalidate the results for that subject. With the group FIAC data we find one subject (and possibly two more) that demonstrate a different pattern of activity. None of our conclusions would be arrived at by simply looking at images of t statistics, demonstrating the importance of model assessment through exploration of the data and diagnosis of model assumptions. *Hum Brain Mapp 27:442–451, 2006.*

Key words: massively univariate modeling; diagnosis; fMRI; FIAC; SPMd

INTRODUCTION

The standard approach to modeling functional MRI (fMRI) data is a linear regression model, with one model for each voxel. This so-called massively univariate approach embodies a series of assumptions at each voxel. Specifically, a linear model assumes (1) mean zero errors, which implies that the model is "correct" and does not lack any important predictors; (2) constant error variance, in particular, that the magnitude of the errors does not vary systematically; (3) independent errors or errors that follow a specified dependence structure; and (4) errors that follow a normal distribution. The first three assumptions are required to ensure that the estimates obtained (e.g., percent BOLD change) are optimal, in that they are unbiased and have the minimum possible variance. The last assumption is required for the calculation of P values, which determine the statistical significance of the estimated effects. When any of these four assumptions are in question, the validity of the final results are unsure, as there might be an increased rate of false-positives; just as troubling, there may be increased false-negatives, that is, reduced power.

For a single univariate dataset, the diagnosis of linear model assumptions is straightforward and routine: The residuals (data minus fitted value) are plotted in various ways to identify any structure (to check assumption 1), to see if the residuals' variance changes systematically (assumption 2), whether the residuals exhibit autocorrelation (assumption 3), and whether the residuals follow a normal distribution (assumption 4). For fMRI, when fitting 100,000 linear models, however, it is impractical to examine 100,000 sets of residual plots. To address

Contract grant sponsors: National Institute of Mental Health, National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering.

*Correspondence to: T.E. Nichols, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: nichols@umich.edu

Received for publication 28 October 2005; Accepted 27 January 2006
DOI: 10.1002/hbm.20253
Published online in Wiley InterScience (www.interscience.wiley.com).

TABLE I. Diagnostic statistics used to create Model Summary images

Diagnostic statistic/Model Summary	Assumption assessed	Null distribution
Cook-Weisberg	Homogeneous Var. $\mathrm{Var}(\varepsilon_i) = \sigma^2$	Chi-squared
Durbin-Watson	Zero autocorrelation $\mathrm{Cov}(\varepsilon_i,\varepsilon_j) = 0$	Beta
Cumulative periodogram	Independence Var(ε) $= \sigma^2 I$	Kolmogorov-Smirnov (tabulated)
Shapiro-Wilk	Normality $\varepsilon \sim$ Normal	
Outlier count (over scans)	Homogeneous errors artifacts	Binomial
Standard deviation	Artifacts	

These statistics are used to create images, where each voxel assesses the validity of a particular assumption.

this problem we have created a set of tools to critically assess the linear model assumptions of fMRI [Luo and Nichols, 2003], and implemented these ideas in a toolbox for SPM2 (Wellcome Department of Imaging Neuroscience, London, UK), Statistical Parametric Mapping Diagnosis (SPMd, Department of Biostatistics, University of Michigan, Ann Arbor, http://www.sph.umich.edu/ni-stat/SPMd/).

The general strategy of SPMd is to create images and time series (or "subject series," for group data) that summarize evidence of assumption violations. The summary data is then interactively explored to identify voxels and scans with possible problems, and then view the "detail," the model fit at individual voxels, or the residual images for specific timepoints (subjects).

The purpose of this work is to demonstrate the SPMd toolbox with the publicly available Functional Image Analysis Contest (FIAC, http://www.madic.org/fiac) data, showing how to assess violations of the assumptions on an individual subject and group analysis. We also report on two new but minor modifications to our method which account for autocorrelation and the particulars of group-level data.

MATERIALS AND METHODS

For each voxel of fMRI data, a general linear regression model (GLM) is fitted by:

$$Y = X\beta + \varepsilon \quad (1)$$

where Y is an N vector response, X is an $N \times p$ design matrix of predictors, β is a p vector of unknown parameter, and ε is an N vector for unknown, random errors. Traditionally, linear model errors are assumed to be identically, independently and normally distributed, $\varepsilon \sim N(0,\sigma^2 I)$ (where "$\sim$" denotes "is distributed as"), and which leads to the ordinary least squares (OLS) estimates $\hat{\beta}=(X^T X)^{-1} X'Y$ and residuals $e = Y - X\hat{\beta}$. While fMRI data are known to be temporally correlated [Friston et al., 2003], in our initial work [Luo and Nichols, 2003] we used a working assumption of independence and sought to detect traditional violations of linear modeling assumptions.

Whereas traditional linear model diagnosis consists of examining plots of residuals, we use diagnostic statistics which are functions of the residuals. Table I shows the diagnostic statistics used and the assumptions they assess [see Luo and Nichols, 2003, for details]; we call these measures Model Summaries, as these statistics summarize the quality of the model fit at each voxel. Where a null distribution is available, we transform the statistic into a $-\log_{10} P$ value image so that all statistic measures have a common scale. For example, the Cook-Weisberg tests for heterogeneous variance by regressing the squared residuals on a potential explanatory variable; if the regression is significant, it supplies evidence that the variance varies with the explanatory variable and that the errors do not all have the same variance.

While Model Summaries are images that assess model fit at each voxel, Scan Summaries (Table II) are N vectors (time series for intrasubject fMRI data, subject-series for group data) where each element assesses problems over an entire image. The two other components of our method are Model Detail and Scan Detail. Model Detail consists of traditional diagnostic plots, showing residual and data fit plots for a single voxel. Scan Detail is simply the standardized residual images, viewed in series, so as to localize the spatial and temporal extent of an artifact.

In Luo and Nichols [2003] we recommend starting with Scan Summaries, to identify any problem images, and then viewing Model Summaries, to localize problem voxels; then for each voxel with possible problems, using the Model Detail to assess the traditional model assumptions and the goodness-of-fit; lastly, if a particular scan appears to have an artifact, the Scan Detail can be used to find exactly what portion of the brain is affected by the artifact. Taken together, these four views of the data can be used to efficiently

TABLE II. Diagnostic measures used to create scan summary time series

Scan summary	Interpretation
Global intensity	Whole-brain signals or artifacts
Outlier count (over voxels)	Shot noise, artifacts
Preprocessing parameters, e.g., head motion	Suggests cause of artifacts
Experimental predictors	For investigating mismodeled signal in residuals
Averaged residual periodogram	Whiteness of residuals, spectral content of physiological/unmodeled variation

Since no explicit model is fit over the image, there are no formal diagnostic statistics available. Rather, these measures are heuristics that are sensitive to artifacts that corrupt part or all of an image.

assess the assumptions of the model and understand any possible violations.

In this article we extend our previous work in two ways: We use models with temporal autocorrelation ($V \neq I$) in single-subject fMRI datasets, and we apply the method to multisubject, group level data. To account for autocorrelated errors, we let, $W = \hat{V}^{-1/2}$ where $\hat{V}$ is an estimator of the autocorrelation V. We can then "pre-whiten" data (Y) and fit ($X\beta$) by pre-multiplying by W on both sides of Eq. (1), giving $WY = WX\beta + W\varepsilon$. We write this whitened model as:

$$Y^* = X^*\beta + \varepsilon^* \quad (2)$$

where $\varepsilon^* \sim N(0,\sigma^2 WVW^T)$. If $\hat{V}$ is an accurate estimator of V, then $WVW^T = I$ and ε^* will be independent, and OLS can be applied to the whitened model and optimally precise estimates $\hat{\beta}^*$ will be obtained. This entire estimation method is known as generalized least squares (GLS). Finally, we write the GLS residuals as $e^* = Y^* - X^*\hat{\beta}^*$.

Diagnosis of autocorrelated models proceeds just as with independence models, except that the assumptions are now on the whitened model (Eq. 2) and tested on the whitened residuals e^*. One difficulty with whitened models is that each whitened scan contains information from multiple scans. That is, each element of Y^* is a linear combination of several elements of Y, as per $Y^* = WY$; likewise, each element of e^* is a combination of multiple timepoints. This is a problem, since if there is a single corrupt scan, say Y_i, the artifact will be spread over multiple rows of Y^*, expanding the impact of the corrupt scan and making the detection of the artifact more difficult. Our solution for this is to examine both the fit of the whitened model and the original (unwhitened) model using the whitened estimates. That is, in addition to plotting Y^*, $X^*\hat{\beta}^*$ and e^*, we plot Y, $X\hat{\beta}^*$ and $e^{(*)} = Y - X\hat{\beta}^*$; in this way we use the optimal whitened estimates $\hat{\beta}^*$ but observe the fit in terms of the original data units.

Diagnosis of multisubject group models follows in the same general framework as above (reviewing Scan Summaries, then Model Summaries, etc.) with three differences. First, since observations are now over subjects, there is no plausible explanation for interscan correlation, and so the autocorrelation and dependence diagnostics are not used. The most useful diagnostic measures are instead simple outlier counts and normality diagnostics, which can detect unusual subjects. The second difference concerns misregistration of each subject's results. While motion correction is an issue with intrasubject analyses, a much more serious concern is the success of the intersubject registration (a.k.a. spatial normalization). Additionally, we should worry about remaining differences in *functional* anatomy. For example, individual subjects could each exhibit robust results, but if the loci of their activations do not align over subjects, there will be no positive result, or one may obtain a misleading positive result if one only examines a t statistic image. A way to address these concerns is to simply *look* at the data, which leads to the third difference between single-subject and group fMRI analyses diagnosis, the possibility of direct visualization.

TABLE III. Landmarks for assessing intersubject registration

Location/view	Description
Cortical surface, using all orthogonal views	Rostralmost frontal cortex. Anterior surface of occipital pole. Ventralmost extent of temporal pole. Dorsalmost cortical surface. Lateralmost left and right cortical surface.
Mid-sagittal plane and coronal sections	Anterior-, superior-, and posteriormost extent of corpus callosum. Border between cerebellum and occipital cortex. Inferior surface of medial orbitofrontal cortex.
Axial slices	Ventricle surfaces, check for gross size differences.

Unlike a single-subject fMRI analysis, it is in fact practical to simultaneously visualize the entire dataset in 2-D sections, due to the (unfortunately) typical small group sizes. This allows for assessment of intersubject variation in structural as well as functional anatomy. For anatomical images, we systematically check the alignment between individuals in the group and between the group and the atlas at the landmarks described in Table III. These landmarks are chosen as they are easy to identify and will be sensitive to misregistration; certainly, if particular regions are of key interest they should be examined on each subject's anatomical image. On functional images, typically consisting of one difference (or contrast) image per subject, we first check alignment with anatomy and then explore the functional data. We explore the mean functional data noting where voxels are lost due to susceptibility artifacts. In the difference images we look for gross patterns that distinguish one or more individuals; if the principal activation can be clearly seen in each subject, we characterize any variability in location of the activation over subjects.

Instead of going into extensive software details (e.g., which button to click), we instead focus on the general heuristics implemented in the SPMd toolbox. The only operational details of SPMd that are essential are: (1) it is Matlab-based on an extension to SPM; (2) it requires a completed SPM analysis directory in which to operate; and (3) it consists of two separate steps, a "Compute" stage where scan summaries, model summaries and scan detail are precomputed, and a "Visualization" stage, where the different summaries and detail data are interrogated.

Data

We use the FIAC single-subject's block-design data (Subject 3, "fonc3" and "fonc4") and the group dataset [Dehaene-Lambertz et al., 2006]. For the group data we use the summary data for 15 subjects for the difference of the DSt-DSp

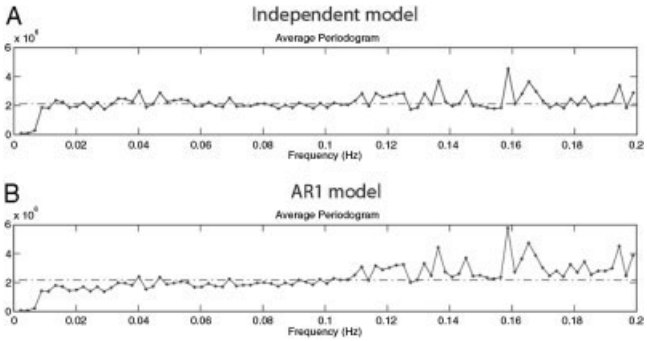

Figure 1.

Average periodogram for a single subject (fiac3, fonc4), for a model assuming independence (**A**) and assuming an approximate AR1 model (**B**). These periodograms indicate that the spectrum of the original data is nearly flat, and the effect of the AR1's whitening is to actually induce excess high-frequency variation.

(Different Sentences – Different Speaker) and SSt-SSp (Same Sentence – Same Speaker) effects.

All data are analyzed with SPM2 (last patch dated July 26, 2005). We carefully describe some methodological aspects of SPM that we have found poorly documented. Single-subject data are realigned to the first scan analyzed. A brain mask is found by retaining all voxels with intensity greater than 80% of a working intracerebral mean image intensity, where the working mean is determined using all voxels having intensity greater than 12.5% of the total volume mean. For intrasubject fMRI modeling, low-frequency noise is deterministically modeled with a Discrete Cosine Transform (DCT) basis (0.0078-Hz cutoff, 7 predictors). Additionally, nonwhite noise is stochastically modeled using a global autocorrelation model based on a first-order Taylor series approximation to an AR(1) model expanded about $\rho_0 = 0.2$, $Cor(\varepsilon_i,\varepsilon_{i+k}) \approx \rho_0^{|k|} + |k|\rho_0^{|k|-1}(\rho_0 - \rho)$. Specifically, the covariance of ε is taken to be a linear combination of matrices corresponding to $\rho_0^{|k|}$ and $|k|\rho_0^{|k|-1}$ (Q_1 and Q_2, respectively), with coefficients estimated with ReML; the ReML estimates are based on the unconstrained N by N covariance matrix of the raw data Y, for the subset of voxels whose OLS F statistic for effects of interest is significant at 0.001 [Friston et al., 2002]. The resulting covariance estimate is normalized into a correlation matrix V, and is then used with GLS at each voxel.

RESULTS

In both the single-subject and the group analysis, we progress from scan summaries, to model summaries, to model detail, and, finally, to scan detail.

Single-Subject Analysis

For the single-subject data, we first review the global aspects of the analysis, specifically the autocorrelation modeling. For the two runs considered there are roughly the same number of voxels in mask (26,578 and 26,246 voxels for "fonc3" and "fonc4," respectively) with a ~0.96-L search volume; for each run, ~3% of the in-mask voxels contributed to the autocorrelation estimation (840 and 1,031 voxels, respectively).

The estimated global ACFs are very similar to $\rho_0^{|k|}$, suggesting that the best global AR1 model was close to an AR(1) with ρ around 0.2. Another view on the autocorrelation is through the global periodograms, the residual power spectrum obtained by averaging over all voxels' spectrums. Figure 1 shows two average periodograms, one for an analysis with no whitening and one with whitening with the approximate AR model. (The zero power at low frequencies is due to the DCT basis eliminating those frequencies from the residuals.) For the analysis without whitening (independence model), a relatively flat spectrum is seen, although with some high-frequency spikes. In contrast, the perio-

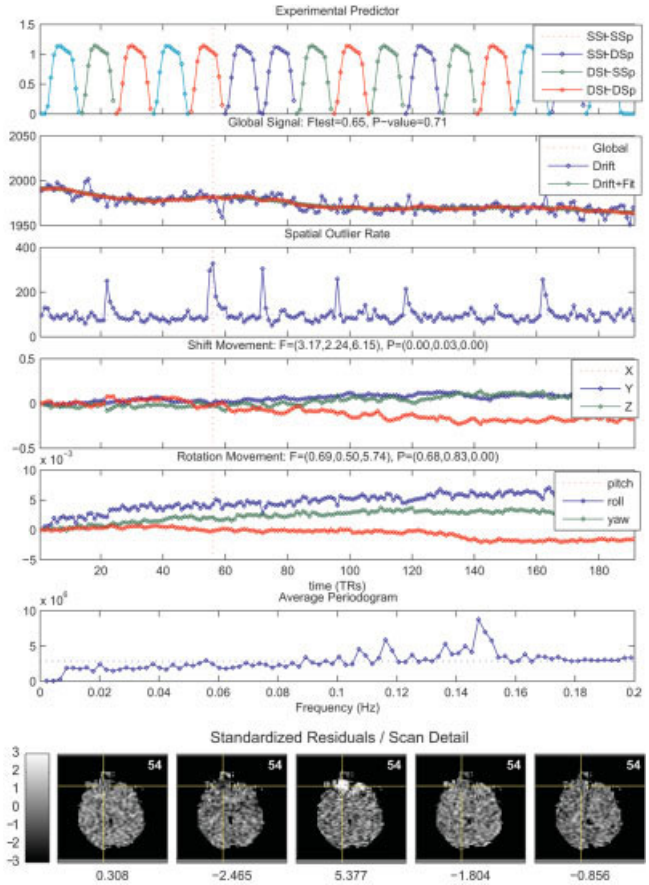

Figure 2.

(**A**) Scan summaries for the Fonc4 data, with the temporal cursor on scan 56. Note that scan 56 has 300% more outliers than would be expected by chance, and following that scan there is a decrease in the global intensity. (**B**) Scan detail (standardized residual images) for images 54–58; the value under each image is the standardized residual under the crosshair. These images show the source of the artifact, a hyperintensity in the orbitofrontal region, just above an air–tissue interface.

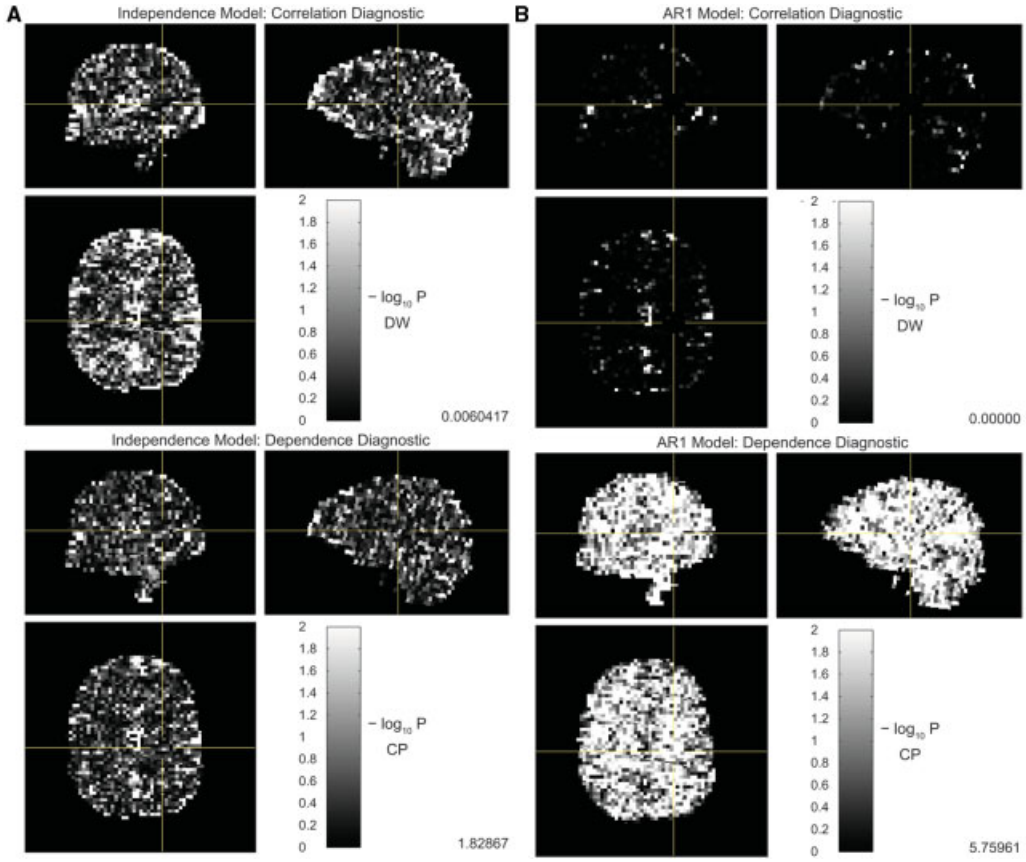

Figure 3.

Scan summaries for assessing noise correlation. (**A**) The CORR (top) and the DEP (bottom) diagnostics for a model assuming independent noise; both have many voxels with values exceeding 2 (i.e., *P* values 10^{-2}), indicating the presence of autocorrelation. (**B**) The CORR (top) and DEP (bottom) diagnostics for the AR1 model; with AR1 autocorrelation modeling the CORR diagnostic has improved, but now the DEP diagnostic is much worse. See Figure 2 for an explanation of this effect.

dograms of the analysis with whitening (AR model) show relative suppression of low frequencies and amplification of high frequencies. Below we comment on the success of the autocorrelation model after considering other diagnostics.

The scan summaries (Fig. 2, top) provide information on problems on a scan-by-scan basis. The plots show no dramatic problems, except that regressions of movement parameters on the experimental design are significant, indicating stimulus correlated motion. For example, in fonc4 all movement parameters except pitch and roll have significant *F* statistics. However, visual inspections of the movement time series indicate that the problem is not severe. The plot of expected outlier counts is typically near 100%, indicating a nominal rate of outliers, although a few spikes are prominent; we revisit these below.

Next we review the model summaries, starting with the correlation diagnostics. Figure 3 shows $-\log_{10} P$ value images of the Durbin-Watson (Zero Autocorrelation, "CORR" test) and Cumulative Periodogram (Independence, "DEP" test) test statistics, for both an independence model and the approximate AR1 model. For the independence model, both the CORR and DEP test image appear somewhat "bright," indicating evidence of correlation (e.g., for fonc3, CORR and DEP identified correlation at 22% and 16% of brain voxels, respectively, when only a nominal 5% rate is expected). With the AR1 model, the CORR image is "dark" and the DEP image is even brighter (for fonc3, CORR only identifies 3.8% problem voxels, while DEP identifies 42% of all voxels, when a nominal 5% rate is expected). Thus, while the CORR diagnostic finds no problem with the AR1 model, the DEP diagnostic actually finds the AR1 model to be worse than the independence model. This matches the findings from the average periodogram in Figure 1.

Before reviewing other Model Summary images, we use the Model Detail view to examine the autocorrelation fit at specific voxels. Figure 4 shows three voxels explored to demonstrate the autocorrelation fit. In the top is shown a voxel with good autocorrelation fit; the lagged residual plot shows autocorrelation in the original data, and after whitening, a nearly spherical distribution of points. The next two rows show examples of autocorrelation overfit, where originally white data now exhibit negative auto-

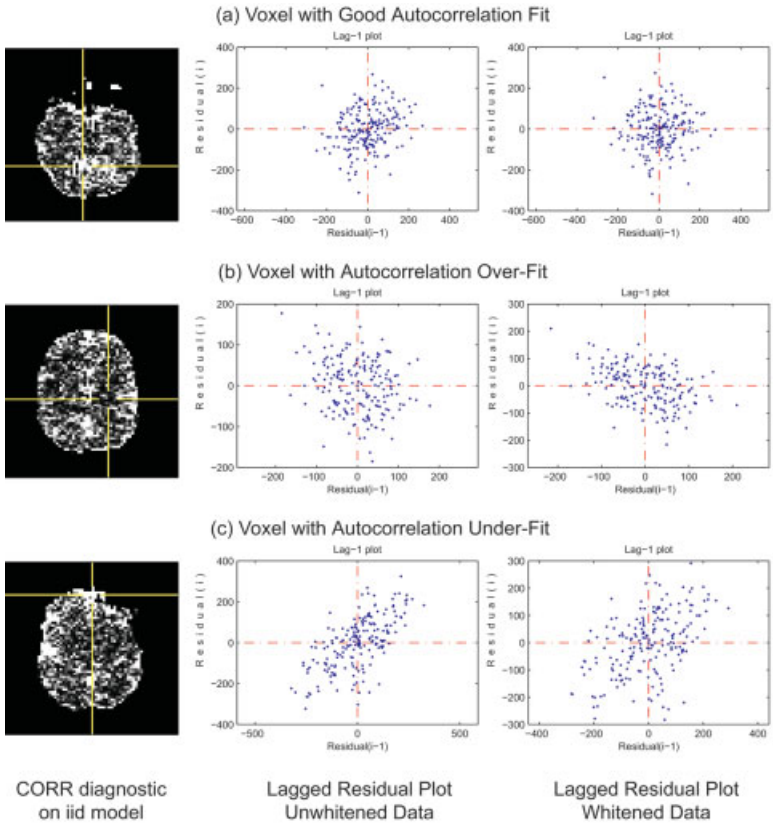

Figure 4.
Exploration of autocorrelation model fit, three examples from different regions. (**A**) Voxel demonstrating good autocorrelation fit. (**B**) Voxel with no autocorrelation initially (note that it appears to be in a white matter area), which exhibits negative autocorrelation after whitening with the approximate AR1 model. (**C**) Voxel with strong autocorrelation that is only partially whitened by the autocorrelation model, and hence still shows positive autocorrelation in the AR1 model.

correlation due to the global whitening, and autocorrelation underfit, where autocorrelation remains after whitening. In total, considering the ACF plots, the average periodograms, the correlation diagnostics and these lagged residual plots, it is apparent that the dependence structure of the data is not well fit by the approximate AR1 model. Specifically, we find that the global model cannot adapt to the local variation in autocorrelation, and that the whitening has overall induced excess high-frequency variation in the residual data.

Having diagnosed the dependence structure of the noise, we turn to traditional assumptions of good model fit (i.e., mean zero errors), homogeneous variance, and normality. Using Model Summary images, we find the typical problems with normality and independence in the brain stem, and medial inferior regions as well as near ventricles and major blood vessels; notably, the standard deviation in these regions is 3 to 10 times higher than typically found in cortex. Such findings are explained by respiratory and cardiac effects and are not much of a concern *unless* the investigator is interested in tissue in or near these regions, in which event physiological effects should be modeled and removed from the data.

The residual standard deviation image identifies an unusual pair of vertical "columns." Figure 5 shows two columns of voxels exhibiting large residual standard deviation, found in both of the sessions. A highly variable voxel in the center of the reconstructed slice is not uncommon, and reflects poor normalization of *k*-space data. Unusual, however,

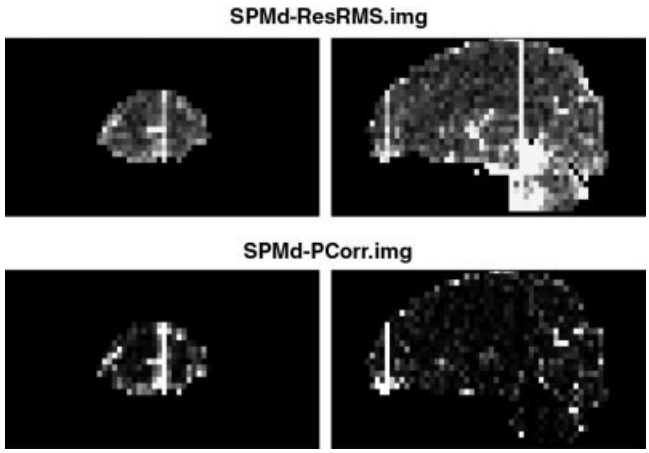

Figure 5.
A comparison of standard deviation image and CORR diagnostic image. The standard deviation image shows two vertical bars, indicating regions of increased variability. The posterior bar resembles an artifact due to poor normalization of the *k*-space data before reconstruction; note that there no evidence of autocorrelation for that bar. The anterior bar is unusual in that, in addition to high variance, it also contains significant evidence of autocorrelation. Inspection of voxel time series finds a consistent 43-second cycle variation (0.023 Hz) that is unique to voxels in that column of voxels. (Note: this display was not oriented into approximate Talairach space, so as to clearly show the artifact.)

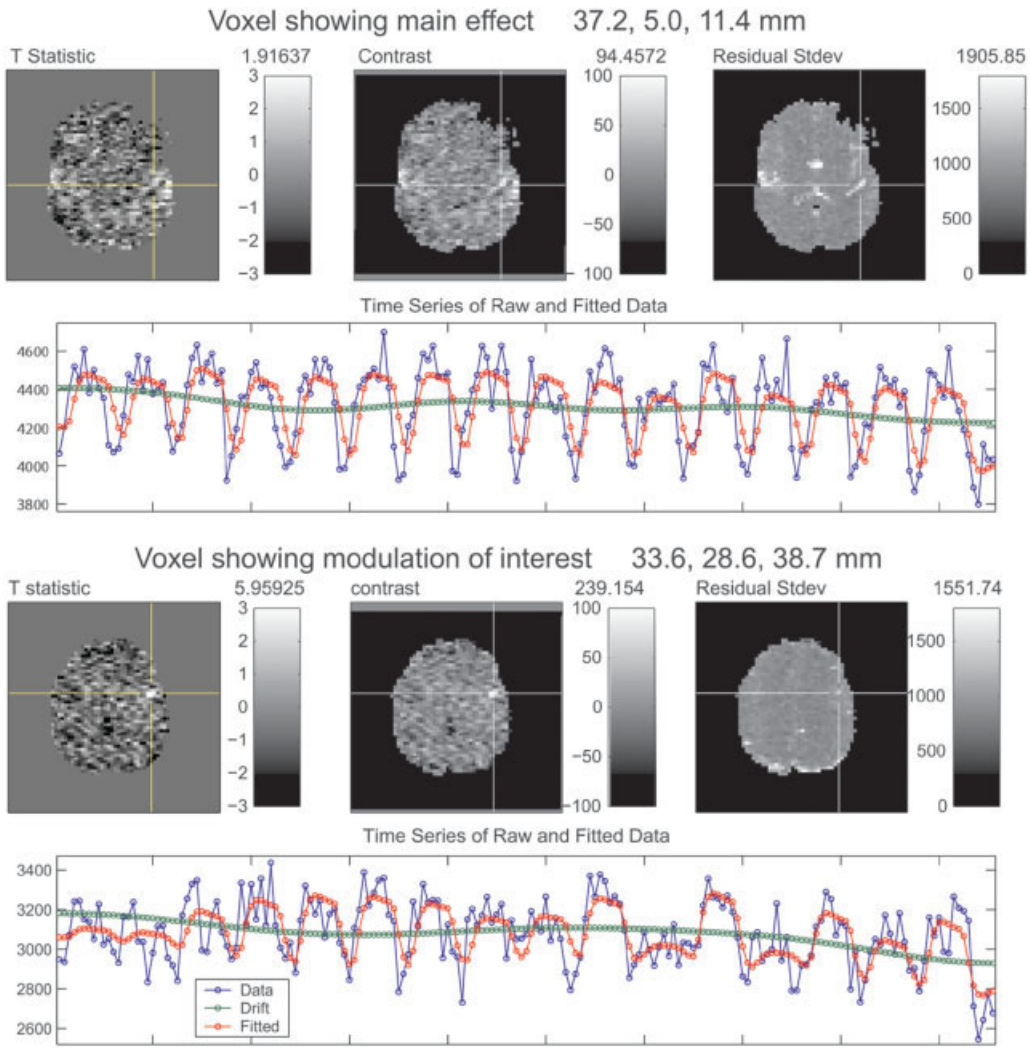

Figure 6.

Voxels showing main (sentence) and differential (DSpDSt-SSpSSt) effects of interest. Top: Voxel in primary auditory cortex, exhibiting the main effect of the sentence presentation (note: the T and Contrast images are for the differential effect, and so are not exceptional). Bottom: Voxel in prefrontal cortex showing the "same" vs. "different" effect. Note that in both voxels the residual standard deviation is relatively high, attributable to lack of fit seen in the time series; specifically, the data seems to rise and fall earlier than the model.

is the second, anterior column that exhibited colored noise, as noted in the CORR image; examination of Model Detail for voxels in the anterior column revealed a prominent 43-second (0.023 Hz) pattern of unknown origin.

Successful activation is noted in the datasets, although active voxels tend to have slightly greater-than-usual residual standard deviation and autocorrelation diagnostics. Figure 6 shows a voxel in the auditory cortex (top) showing the main effect of the stimuli presentation, and in the lateral prefrontal cortex (bottom) exhibiting the DSt-DSp vs. SSt-SSp modulation. The increased standard deviation and CORR diagnostics are due to systematic lack-of-fit in the model, with the unmodeled variation inducing autocorrelation into the residuals. Specifically, it appears that the data rise and fall more quickly than the model, indicating that the canonical HRF used by SPM has greater delay than found in this subject. While including a temporal derivative in this model could have improved model fit, it could not account for the underestimation of the response magnitude, which is clearly seen by the under- and overshoot of the data relative to the model.

Finally, returning to the outlier spikes noted above in Scan Summary, we use Scan Detail to determine the source of increased outliers around scan 56. Figure 2, bottom, shows standardized residuals for scans 54–58, and an oribitofrontal artifact is clearly observed in scan 56. Scan 56 exhibits a hyperintensity, while subsequent scans appear relatively dark in that region. The cause of this is possibly movement-related,

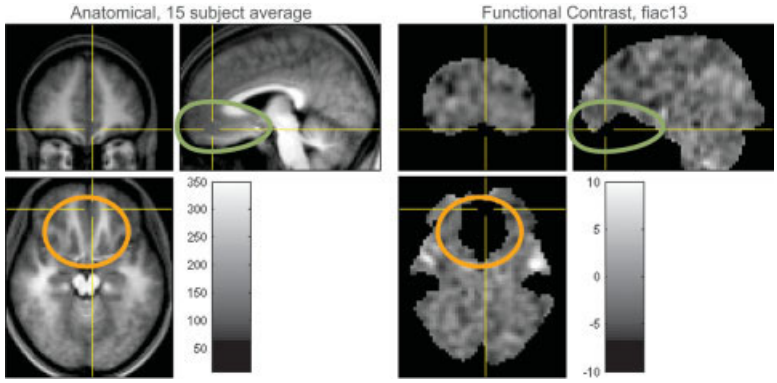

Figure 7.
Visualization of susceptibility voids through comparison of a typical functional (contrast) image with the averaged T1 anatomical image. Large extents of the orbital frontal and nearby cortices are missing.

although no dramatic movements correspond to these spikes. While the affected scans could be removed from the analysis, a more pragmatic approach is simply to note the region and be skeptical of activations that might appear there.

In summary, we have identified several problems with the data and model, most notably with the autocorrelation model. Since the autocorrelation model is principally concerned with obtaining accurate *P* values (and secondarily, with optimally precise estimates) the implications for this problem are most severe when inference is made on an individual subject. If the interest is in the group analysis, then the quality of the autocorrelation model is less of a concern.

Group Data Analysis

The group data analysis is based on 15 subjects' contrast images, for the "different vs. same" contrast. There are 45,494 voxels in the mask, a 1.23-L search volume. We check the extent of susceptibility artifacts using the contrast images (ideally, mean functional data for each subject, in atlas space, would be used), noting the location of the edge of the analysis mask. Figure 7 shows the extensive susceptibility voids in the orbitofrontal cortex, common with echo-planar imaging (EPI) acquisitions. Since the hypothesized regions of response lie outside of this region, we are not very concerned with this finding.

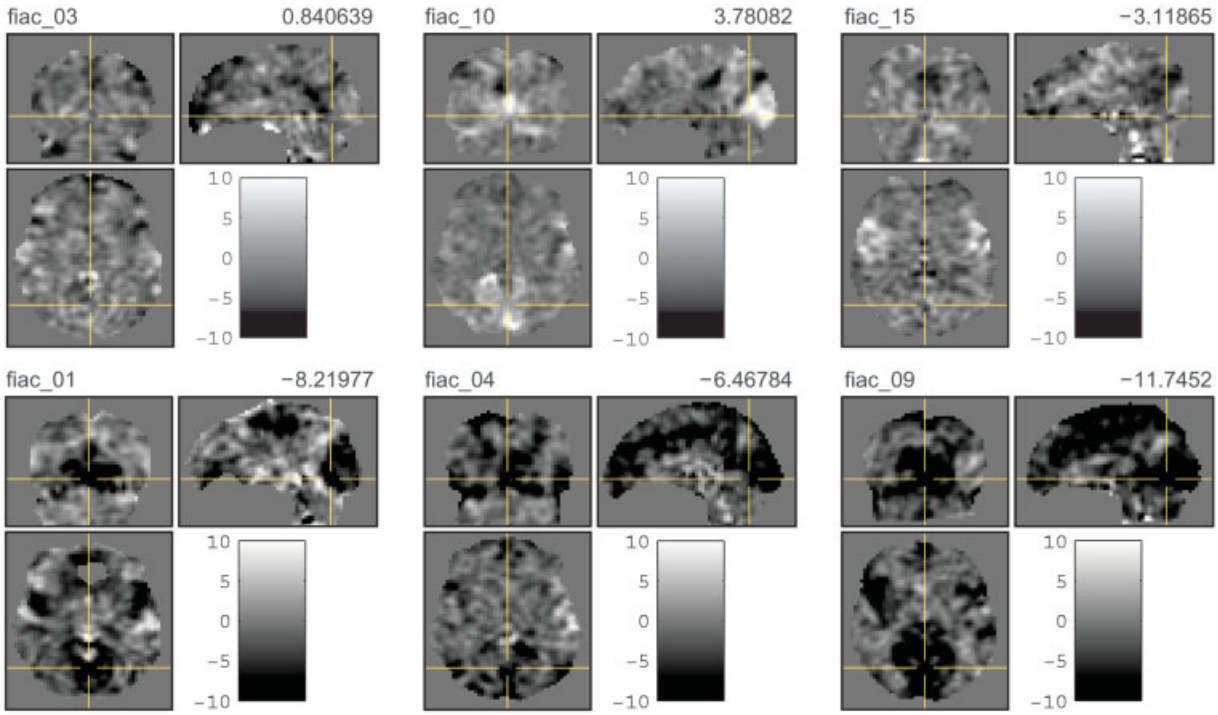

Figure 8.
Use of orthogonal viewers to view the group dataset simultaneously (due to space limitations, only six of the subjects are shown). Most subjects appeared "gray," meaning their contrast data lies near zero (see top row, FIAC Subjects 3, 10, and 15). In distinction, Subject fiac9 (bottom row, right) shows extensive decreases in visual cortex, anterior cingulate, and insular cortices; Subjects fiac2 and fiac4 also show similar patterns.

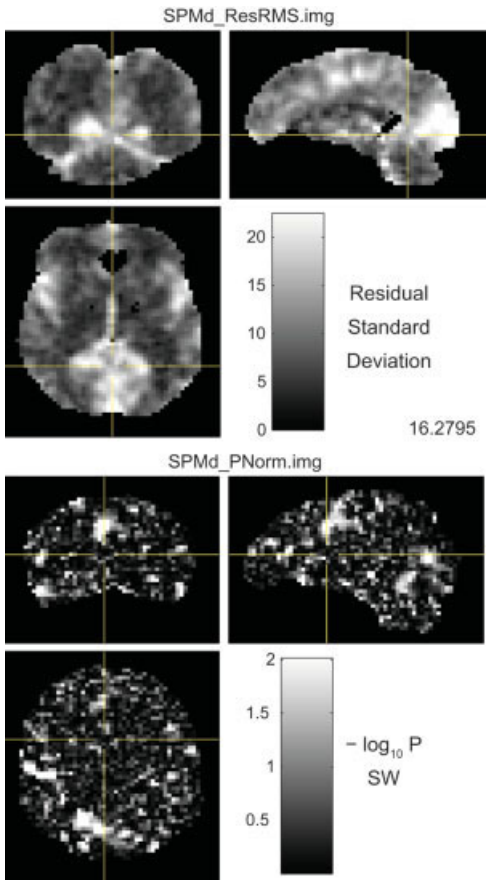

Figure 9.
Useful scan summaries for the group data, residual standard deviation, and the Shapiro-Wilk normality diagnostic. (**A**) The standard deviation image shows considerable structure; the primary visual cortex is particularly well-defined in the axial view, indicating visual activity was quite heterogeneous over subjects; the insular and superior midline regions (possibly SMA) were also quite variable. (**B**) The normality diagnostic was mostly small (dark) throughout most of the brain, except in a handful of regions; these three orthogonal views show evidence of nonnormality in extrastriate regions, a left superior temporal region, and superior midline region.

Before examining Scan Summaries, we visually explore the entire 15-subject dataset. Using a 5 × 3 array of orthogonal viewers (SPM's CheckReg facility) this is practical, if perhaps requiring a large display. We first explore the anatomical images in the standard space, to ensure that the intersubject registration to the atlas is successful. Results of the alignment check are shown in the Appendix; in general, the alignment is very good, with the only notable lack of alignment being variation in corpus callosum shape and location of the ventral-most extent of the temporal pole. The direct inspection of the functional data is striking (Fig. 8). While most subjects' contrast images are homogeneous and exhibit no particular pattern (i.e., appear mostly gray), Subjects fiac1 and fiac9 exhibit dramatic decreases in the visual and superior midline regions (i.e., are very black); fiac4 also exhibits some degree of decreases along the midline. (Note that in this view it is crucial to use the same grayscale intensity window for all subjects.)

The only scan summary of use for group analysis is the outlier rate (Fig. 9), and we notice that fiac9 has 699% of the number of outliers expected by chance ($\alpha = 0.05$) and Subjects 4 and fiac4 and fiac8 have over twice the expected rate (212% and 219%, respectively). Other subjects have outlier rates far below the nominal 100%, where *low* outlier rates suggest an *inflated* residual standard deviation, attributable to one or more outlier subjects—we revisit this below.

The normality diagnostic and the residual standard deviation in Model Summaries are the most useful in the group analysis (Fig. 9). The normality test is large near the visual cortex and in a superior mid-line region. Model Detail for voxels in these regions reveal that the nonnormality is attributable to 1 or 2 outlier subjects, mostly fiac9, sometimes in conjunction with fiac8 or fiac4.

The *t* statistic image is one type of Model Summary, and Figure 10 shows a potentially significant region. With a *t* of 5.446, uncorrected *P* value of 0.000043, it could be an important activation. However, when inspecting the contrast image, the region is not exceptional (other regions in the con-

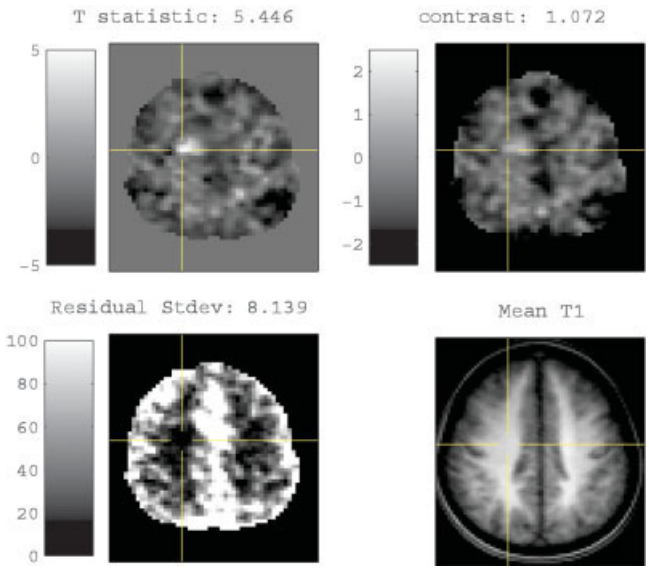

Figure 10.
Probable false-positive region found with Model Summaries. Exploration of Model Summaries identified a prominent region in the *T* statistic image (top left), with a peak *t* of 5.446, uncorrected *P* value of 0.000043. The region is found to be unexceptional in the mean effect or contrast image (top right), but is found to be a region with exceptionally *low* variance (bottom left). The region is most surely a false-positive region, as inspection of the mean anatomical shows the voxel to be squarely within white matter (bottom right), underscoring the need for corrected significances.

trast image with similar magnitude do not appear bright in the *t* image), and only when the standard deviation image is viewed can it be understood that this "activation" speaks more to a local decrease in variability than a prominent increase response magnitude; moreover, the corresponding anatomical image shows the region is squarely in white matter. Interestingly, that voxel is not significant when correcting for multiple comparisons (corrected *P* value of 0.9878 and 0.1122 for FWE and FDR, respectively), suggesting the data there are consistent with the null hypothesis.

Lastly, the Scan Detail (standardized residual images) shows that Subject fiac9 is generally very "dark," with almost all of its residuals negative (it was found that 81.5% of that fiac9's voxels are less than zero, far from the nominal 50%). Subjects' fiac1 and fiac4 appear dim and also have a large proportion of negative residuals.

In short, we find that the FIAC group data demonstrates substantial intersubject heterogeneity, with at least one and likely three subjects (fiac9, fiac2, fiac4) behaving in a fundamentally different manner from the other subjects. Notably, none of these subjects were reported as having slept or having trouble with the task. A comprehensive analysis would consider excluding these subjects, or, perhaps better, using robust methods [Wager et al., 2005] that can implicitly down-weight unusual subjects.

DISCUSSION AND CONCLUSION

We have presented two new aspects of diagnosis of linear models for fMRI, the evaluation of correlation models for single-subject fMRI, and of group models for multisubject data. We have found several anomalies in the single-subject data, but none particularly catastrophic. In contrast, for the group analysis there are a number of measures that suggest that Subject fiac9 is completing the task in a fundamentally different manner than other subjects; Subjects fiac1, fiac4, and fiac8 also appear to diverge from the group in different aspects. All of the diagnosis methods presented here are implemented in SPMd, a toolbox for SPM. SPMd provides dynamic graphical tools that make it possible to explore the large datasets common in fMRI. For the group analysis, however, most of the diagnosis consist simply of direct exploration of all subjects' contrast images simultaneously with SPM's Check-Reg facility; hence, with small group analyses, exploration is straightforward and possible with any software supporting multiple orthogonal viewers.

To conclude, especially in light of recent challenges to the validity of fMRI analyses [Dobbs, 2005], we stress the importance of thoroughly exploring one's data, and carefully assessing the validity of a model's assumptions. The end result will be inferences that are trustworthy and a greater understanding of the limitations of one's data.

REFERENCES

Dehaene-Lambertz G, Dehaene S, Anton JL, Campagne A, Ciuciu P, Dehaene GP, Denghien I, Jobert A, LeBihan D, Sigman M, Pallier C, Poline JB (2006): Functional segregation of cortical language areas by sentence repetition. Hum Brain Mapp 27:360–371.

Dobbs D (2005): "Fact or Phrenology?" Sci Am Mind 1:24–31.

Friston KJ, Frith CD, Dolan RJ (2003): Human brain function. New York: Academic Press.

Friston KJ, Glaser DE, Henson RNA, Kiebel S, Phillips C, Ashburner J (2002): Classical and Bayesian inference in neuroimaging: applications. Neuroimage 16:484–512.

Luo W-L, Nichols T (2003): Diagnosis and exploration of massively univariate neuroimaging models. Neuroimage 19:1014–1032.

Wager T, Keller M, Lacey S, Jonides J (2005): Increased sensitivity in neuroimaging analyses using robust regression. Neuroimage 26:99–113.

APPENDIX

Below is the assessment of the intersubject registration, as noted at the landmarks described in Table III. For each landmark, first the degree of agreement within the group is given ("within"), followed by the agreement of the group mean T1 with the avg152.mnc MNI atlas ("atlas"). If a landmark or within/atlas designation is omitted, the relevant alignment is satisfactory. The following shorthand notations are used: A/P = Anterior-Posterior; S/I = Superior/Inferior, CC = Corpus Callosum and Ss = subjects.

Location/view	Description
Cortical surface, using all orthogonal views	Rostral-most frontal cortex. Within: fiac2 relatively posterior.
	Anterior surface of occipital pole. Within: Some Ss have asymmetric pole.
	Ventral-most extent of temporal pole. Within: Substantial S/I variability and image quality poor.
	Lateral-most left and right cortical surface. Within: fiac0 more medial on right side, and fiac3 and fiac8 are more medial on the left side. Atlas: Mean T1 is more lateral than the atlas on left side.
Mid-sagittal plane	Anterior-, superior- and posterior-most extent of corpus callosum. Within: Superior point of CC is most variable of CC landmarks; substantial variability in CC shape. Atlas: Mean T1 anterior CC displaced 4 mm anteriorly and superiorly relative to atlas; mean T1 superior CC posterior (7.5 mm).
Mid-sagittal plane and coronal sections	Inferior surface of medial orbitofrontal cortex. Within group: Substantial S/I variation, up to 10 mm difference between Ss.
Axial slices	Ventricle surfaces, check for gross size differences. Within group: Overall good consistency, exceptions include fiac10, with relatively large anterior horn of lateral ventricle, and fiac15, with relatively large posterior horn of lateral ventricle. Atlas: Overall good alignment, except on the mean T1 the posterior horns of lateral ventricle were more lateral than the atlas's ventricles, more so on left than right.