Invited Review

Report on a Multicenter fMRI Quality Assurance Protocol

Lee Friedman, PhD[1]* and Gary H. Glover, PhD[2]

Temporal stability during an fMRI acquisition is very important because the blood oxygen level-dependent (BOLD) effects of interest are only a few percent in magnitude. Also, studies involving the collection of groups of subjects over time require stable scanner performance over days, weeks, months, and even years. We describe a protocol designed by one of the authors that has been tested for several years within the context of a large, multicenter collaborative fMRI research project (FIRST-BIRN). A full description of the phantom, the quality assurance (QA) protocol, and the several calculations used to measure performance is provided. The results obtained with this protocol at multiple sites over time are presented. These data can be used as benchmarks for other centers involved in fMRI research. Some issues with the various protocol measures are highlighted and discussed, and possible protocol improvements are also suggested. Overall, we expect that other fMRI centers will find this approach to QA useful and this report may facilitate developing a similar QA protocol locally. Based on the findings reported herein, the authors are convinced that monitoring QA in this way will improve the quality of fMRI data.

Key Words: quality assurance; fMRI; phantom; stability; SNR
J. Magn. Reson. Imaging 2006;23:827–839.
© 2006 Wiley-Liss, Inc.

MOST FMRI STUDIES are based on the blood oxygenation-level-dependent (BOLD) contrast (1), and this signal change is a small fraction of the raw signal intensity. According to Matthews (2), BOLD fMRI signal changes in typical tissue voxels (on the order of 3 × 3 × 3 mm) with usual sorts of stimuli are not more than a few percent at 1.5T. To accurately measure such small signal changes, an MR system must have intrinsic image time-series fluctuation levels that are much lower than the expected signal changes. Although there are many general aspects to MR quality control (e.g., geometric accuracy, contrast resolution, ghosting level, spatial uniformity, etc.) (3–5), this paper emphasizes our experience with a quality assurance (QA) protocol focused on scanner performance stability, both within a run and across days or weeks.

As part of the protocol for a multicenter fMRI consortium (FIRST-BIRN, www.nbirn.net) comprised of 12 centers (see Table 1 for a list), a weekly QA assessment was initiated. Although most centers in the consortium had some form of QA protocol in place before the program began, we found that they were often inconsistently applied. Moreover, initial scanner performance was found to be somewhat unstable and to vary widely across sites before a uniform and regular QA program was instituted. Indeed, this may be the case for many imaging facilities engaged in fMRI studies. As will become clear, more frequent and regular assessments can be quite helpful for detecting gradual and acute degradation of scanner performance. Frequent assessments can also be helpful for monitoring software and hardware upgrades. If such changes in performance are not controlled, they can add unexplained variance to data collected over long periods of time, and can be especially harmful for longitudinal studies.

We wish the reader to note that our report focuses entirely on stability, signal-to-noise ratio (SNR), drift, and other hardware performance issues related to MR scanners. We specifically do not address the issue of reproducibility and compatibility of the analysis software between sites, since that important topic is beyond the scope of this review.

MATERIALS AND METHODS

Over the past few years, one of the authors (G.H.G.) has developed and employed a specific protocol for assessing scanner stability. The basic idea is to collect a time-series of images using a phantom and perform an automated analysis of the time-series. In this section the approach will be described in detail. In the next section we will show how this protocol has been used by the FIRST-BIRN consortium to monitor QA.

Phantom

The phantom for the Glover stability QA protocol (GSQAP) is a 17-cm-diameter spherical plastic vessel

[1]Department of Psychiatry, University of California–Irvine, Irvine, California, USA.

[2]Department of Radiology, Stanford University, Stanford, California, USA.

Contract grant sponsor: FIRST Biomedical Informatics Research Network, NCRR/NIH; Contract grant number: 5 MOI RR 000827.

*Address reprint requests to: L.F., 1312 Michael Hughes Dr NE, Albuquerque, NM 87112. E-mail: lfriedman10@comcast.net

Received September 27, 2005; Accepted February 17, 2006.
DOI 10.1002/jmri.20583
Published online 28 April 2006 in Wiley InterScience (www.interscience.wiley.com).

Table 1
Description of Hardware and Sequences of the Nine Sites (10 Scanners) Participating in this Study, Five 1.5 T Sites, Four 3 T Sites, and One 4 T Site

Center	Abbreviation	Field strength	Manufacturer	RF coil type	Functional sequence
Brigham & 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
Massachusetts 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

(Dielectric, Inc., Madison, WI, USA) filled with a doped agar gel. The purpose of the doping is to approximate the T1 and RF conductivity of brain tissue so that the coil loading and NMR equilibrium conditions will be similar to those in a typical fMRI scan protocol. Details of the agar preparation are provided in the Appendix. The gel is preferred to doped water filler because the T2 and magnetization transfer characteristics of agar are more similar to those of the brain. Also, the use of agar gel avoids a long settling time and ensures less influence from vibration.

Scanning Protocol

The goal of the QA procedure is to measure the stability of the scanner under conditions that are as similar as possible to those of a typical fMRI experiment. Thus the same head coil and scanning sequence are employed. The FIRST-BIRN parameters are listed in Table 2. In any case, a protocol with a gradient and RF duty cycle that are at least as strenuous as those used in normal fMRI scans should be employed. The normal reconstruction methods employed in fMRI studies should be utilized, since algorithmic differences can be reflected in the measurements. For example, a correction for spatial drift or distortion due to magnetic field drift may affect the derived stability measures.

Table 2
FIRST-BIRN fMRI Sequence Parameters

Acquisition type	EPI or spiral gradient echo recalled
Scan plane	Humans: axial oblique (AC-PC line) Phantom: straight axial
Field of view	22 cm
Slices	27, 4-mm-thick, 1-mm interslice gap
TR	2000 msec
TE	30 msec (3 T/4 T), 40 msec (1.5 T)
Flip angle	90 degrees
Bandwidth	≥ ± 100 kHz
Matrix	64 × 64
Number of volumes collected	For QA: 200 + optional warmup volumes
Scan time	6.67 minutes

All of the analyses are based on a time-series of 200 images from the middle slice through the phantom. The first two collected volumes are discarded to allow NMR and eddy-current equilibrium to be achieved (some scanners may also scan but not collect additional "warm-up" volumes). The analysis is implemented in Matlab (The MathWorks, Inc., Natick, MA, USA), and several computed images are obtained together with region-of-interest (ROI) analyses of these computed images and the time-series. A representative Matlab script is available at http://www.nbirn.net/resources/downloads.

Signal Image

The signal image is the simple average, voxel by voxel, across the 198 images.

Temporal Fluctuation Noise Image

To calculate the fluctuation noise image, the time-series across the 198 images for each voxel is detrended with a second-order polynomial. The fluctuation noise image is an image of the standard deviation (SD) of the residuals, voxel by voxel, after this detrending step.

Signal-to-Fluctuation-Noise Ratio (SFNR) Image and Summary SFNR Value

The signal image and the temporal fluctuation image are divided voxel by voxel to create the SFNR image. A 21 × 21 voxel ROI, placed in the center of the image, is created. The average SFNR across these 441 voxels is the SFNR summary value.

Static Spatial Noise Image

An extension of a procedure recommended by the National Electrical Manufacturers Association (6) is used to obtain a measure of the spatial noise. The first step is to sum all of the odd-numbered images (sumODD image) and separately sum all of the even-numbered images (sumEVEN image). The difference between the sum of the odd images and the sum of the even images (DIFF = sumODD – sumEVEN) is taken as a raw measure of static spatial noise. If the images in the time-series exhibit no drift in amplitude or geometry, the DIFF image will display no structure from the phantom, and the variance in this image will be a measure of the intrinsic noise.

SNR Summary Value

The static spatial noise variance summary value is the variance of the static spatial noise (DIFF) image across a 21 × 21 voxel ROI centered on the image. The signal summary value is the average of the signal image across this same ROI. Then, SNR = (signal summary value)/√((variance summary value)/198 time points).

Percent Fluctuation and Drift

To compute these summary values, a time-series of the average intensity within a 21 × 21 voxel ROI centered in the image is obtained, as shown in Fig. 1. A second-order polynomial trend is fit to these data (thick smooth line in Fig. 1). The mean signal intensity of the time-series (prior to detrending) and SD of the residuals after subtracting the fit line from the data, are computed. Percent fluctuation equals 100 * (SD of the residuals)/(mean signal intensity). Drift is computed by subtracting the minimum fit value from the maximum fit and dividing by the mean signal intensity. This drift value is also multiplied by 100 to obtain a percentage.

Fourier Analysis of the Residuals

The next analysis in the GSQAP is a Fourier analysis of mean signal intensity in the ROI over time (volume number). After the data are detrended with a second-order polynomial, the residuals are submitted to a fast Fourier transform (FFT). Since the number of volumes is not a power of 2, the FFT is based on a mixed-radix calculation. What is displayed in the GSQAP is the magnitude spectrum with the DC term suppressed (Fig. 2). For the data presented here, the spectrum magnitude values are in units related to the raw signal, and thus they vary from machine to machine. This makes it difficult to apply a universal amplitude criterion for our multisite data. However, the most useful aspect of the spectral plot is to highlight discrete frequencies that can occur because of mechanical vibrations, such as those from the refrigerator ("cold head"), or gradient-induced resonances. An absolute scale is not important for this application. However, future versions of the GSQAP will normalize the spectrum by the raw signal to provide percentage intensities, thus allowing a criterion for acceptable limits to be applied universally.

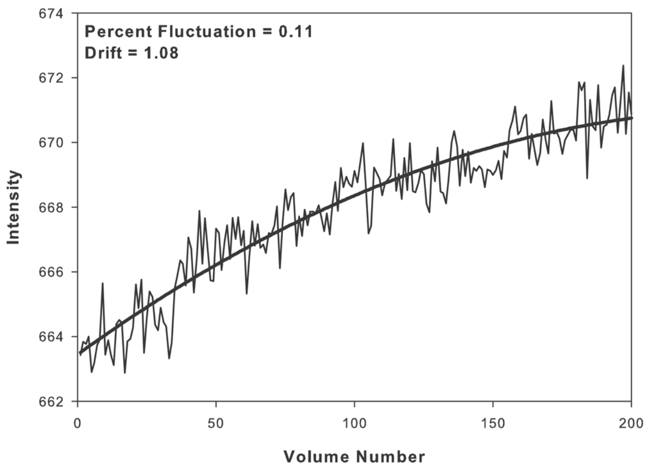

Figure 1. Illustration of percent fluctuation and drift calculations. The abscissa is the volume number (time), and the ordinate is the image intensity. The average intensity within the 21 × 21 pixel ROI placed in the center of the phantom is plotted over time and indicated by the jagged thin line. The thicker line is the fit of a second-order polynomial to the time-series data.

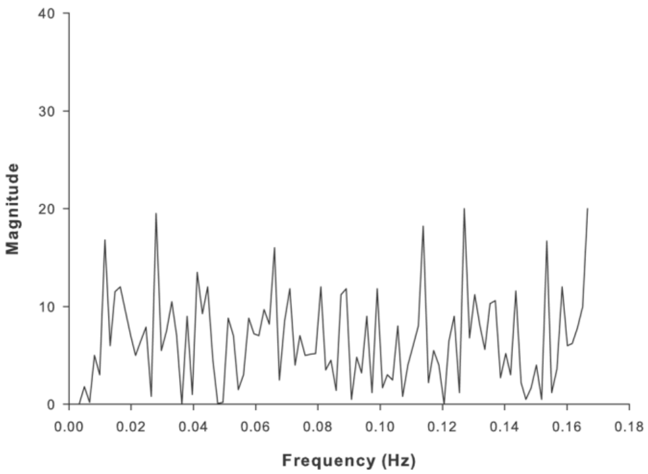

Figure 2. Typical magnitude spectrum from a scanner that does not suffer from a strong periodic noise source. The actual magnitudes of noise are small (<20), and although the spectrum is somewhat spiky, none of the spikes really stand out as abnormally elevated.

One begins analysis of these data with a visual inspection of the magnitude spectrum. If there are no periodic noise sources, the spectrum magnitudes should be low and relatively uniform. In actuality, the spectra are always somewhat spiky, as in Fig. 2, but no individual peaks stand out. (It is possible to compute a statistical probability (*P*-value) for each peak in the magnitude spectrum based on the peak amplitude and an estimate of noise (the two surrounding values) (7).) When there is periodic noise, it shows up as a spike, as in Figs. 3a and 4a. In Fig. 3a a spectrum is presented from the same scanner as in Fig. 2, but note the very large peak (indicated with an asterisk) at about 0.003 Hz. In this case the peak in question was statistically significant ($P = 0.023$). In the lower graph of Fig. 3b, the residual signal and a sine wave representing the time-domain equivalent from the peak in Fig. 3a are shown. Note the obvious presence of this slow oscillation, which was due to a faulty RF-transmit amplifier. Figure 4 presents data from another scanner (UCIR) with a statistically significant ($P = 0.04$) spectral peak at 0.02 Hz.

Weisskoff Analysis

The Weisskoff analysis (8) provides another measure of scanner stability included in the GSQAP. It assumes

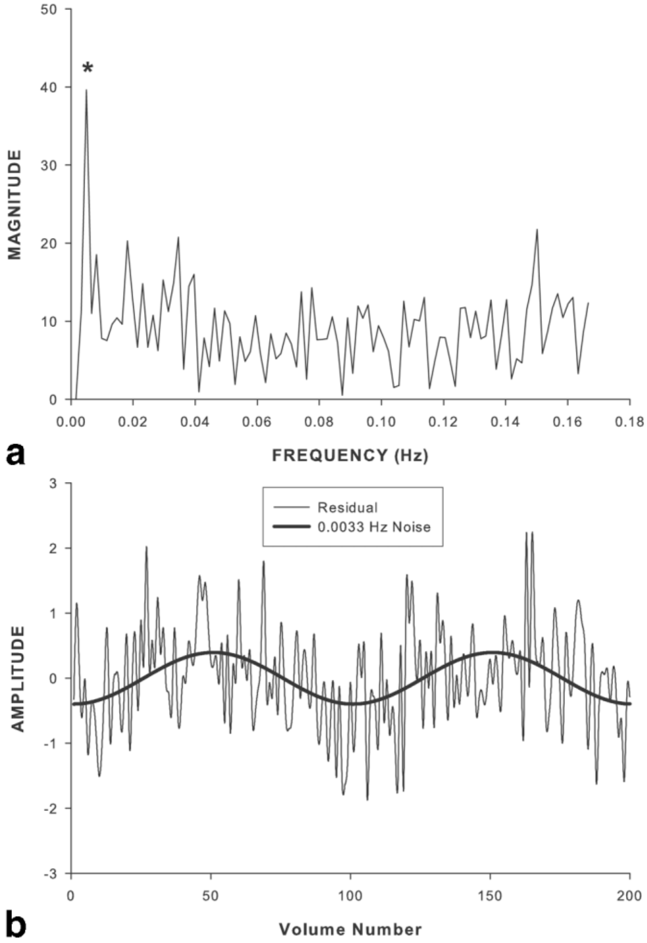

Figure 3. Magnitude spectrum and noise illustration with noise at 0.0033 Hz. **a:** Note in this magnitude spectrum the very noticeable peak at 0.0033 Hz. The asterisk indicates that this peak was statistically significantly elevated compared to its two immediately neighboring frequencies. **b:** Residual time-series signal (after detrending with a second-order polynomial) and the noise component represented by the peak with the asterisk in part a.

that scanner instabilities will impart some increase in the intervoxel correlation, presumably because such instabilities will have some low-spatial-frequency characteristics. If there are no such instabilities (or spatial smoothing in the reconstruction; see below), then each voxel is (relatively) independent of its neighbors, and the coefficient of variation (CV, the SD of a time-series divided by the mean of the time-series) for an ROI should scale inversely with the square root of the number of voxels in the ROI. Thus, for a square N × N voxel ROI, a plot of log(CV) vs. log(N) should follow a declining straight line (8) (Fig. 5a). In practice, as N increases, the reduction in CV plateaus and becomes independent of N (Fig. 5b). This occurs because system instabilities result in low-spatial-frequency image correlations, so that the statistical independence of the voxels is lost. The GSQAP defines a radius of decorrelation (RDC) as CV(1)/CV(Nmax), where Nmax is 21. As shown in Fig. 6, the RDC is the intercept between the theoretical CV(N) and the extrapolation of measured CV(Nmax), where Nmax is 21. Thus the RDC may be thought of as a measure of the size of ROI at which statistical independence of the voxels is lost.

RF-Receiver Gain, RF-Transmit Gain, and Resonant Frequency

When available, the RF-receiver gain, RF-transmit gain, and resonant frequency settings can be quite valuable. Since the phantom and the protocol remain the same, in principle the amplifier gains should remain constant. Changes in these levels can indicate a hardware problem. The resonant frequency can drift over time because the windings of the magnet are not perfect superconductors, and thus there is a small time-dependent drop in the current flowing in the windings and a corresponding downward drift in the magnetic field strength and corresponding resonant frequency, as illustrated in Fig. 7.

RESULTS

Comparing Nine Scanners on Four Key GSQAP Measures

To give a sense of the range of values produced by the GSQAP on various scanners, we averaged the last 10 measures from nine scanners that are included in the

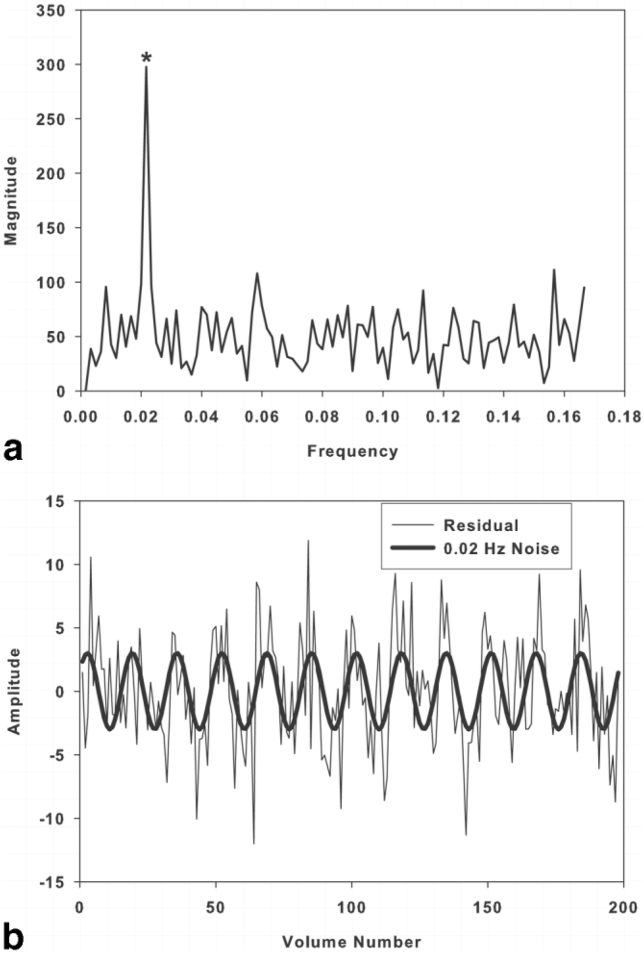

Figure 4. Magnitude spectrum and noise illustration with noise at 0.02 Hz. For additional details see Fig. 3.

FIRST-BIRN project. The scanners are labeled with a four-letter code that is defined in Table 1, which also describes each scanner.

The measures are displayed in Fig. 8. In Fig. 8a the mean SNR (+SD) for the nine scanners is presented, in b the mean SFNR is presented, in c the mean percent fluctuation is presented, and in d the mean drift (absolute value) is presented. The white bars represent 1.5T scanners, the black bars represent 3T scanners, and the striped bar represents a 4T scanner. In general, SNR and SFNR are lower at 1.5T, and at 3T both measures are around 200. The 4T scanner has a markedly higher SNR and SFNR than the other scanners. It appears from a and b that there may be a correlation between SNR and SFNR, and indeed that is the case (r = 0.99; Fig. 9).

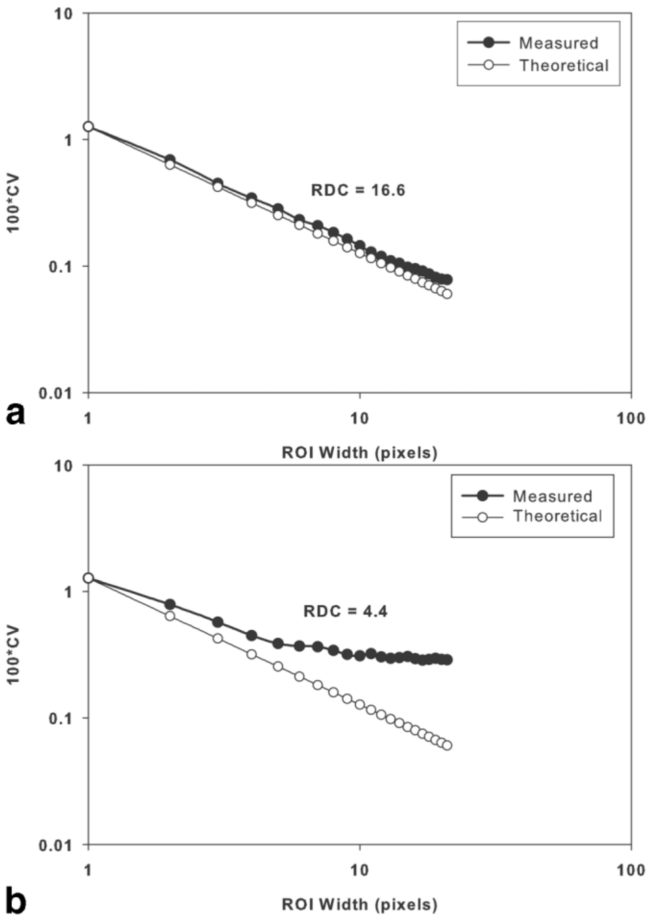

Figure 5. Illustration of the Weisskoff analysis as adapted in the GSQAP. In **a** and **b** the abscissa is the pixel width of a series of ROIs over which the calculations are made (pixel widths = 1-21, ROI dimensions = 1 × 1 to 21 × 21). It is plotted on a log scale. On the ordinate is plotted the CV (SD/mean) *100 for the time-series based on incrementally increasing ROI size. It is also plotted on a log scale. The SD is that of the residuals of the time-series after polynomial detrending. The open circles represent the theoretical expectation assuming that the SD will decline linearly with one dimension of the ROI (the width, which is the square root of (dimension$_x$ × dimension$_y$), for example 21, which is the square root of (21 × 21). The dark filled dots plot the actual data. a: A good Weisskoff plot result (i.e., the data dots are quite close to the theoretical results). b: A poor Weisskoff plot result (i.e., the data deviate from the theoretical results when the ROI width is only 2 pixels, and stop declining at all at an ROI width of approximately 6 pixels.

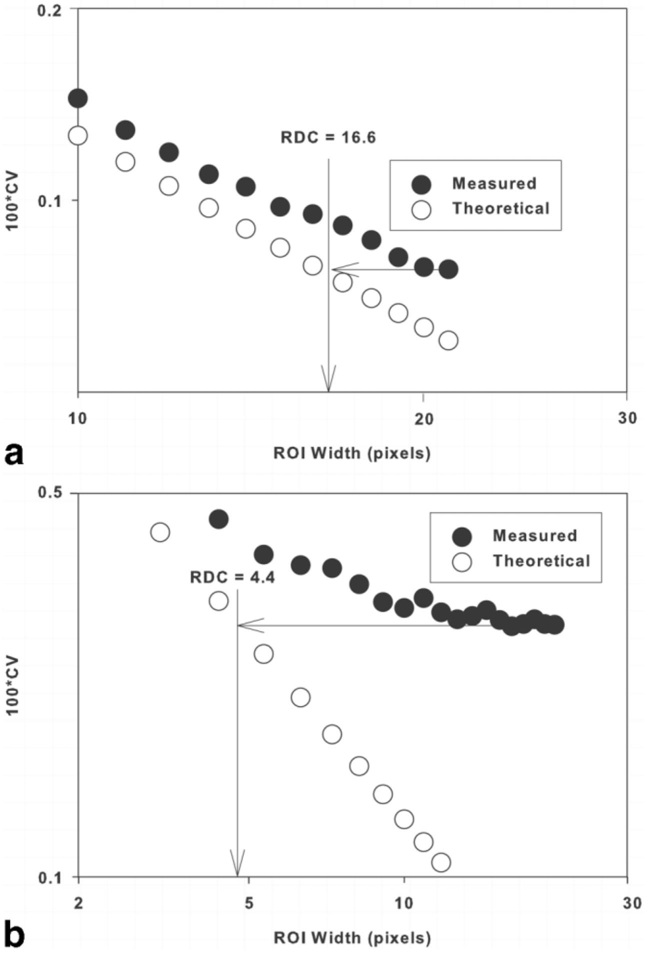

Figure 6. Graphic illustration of the RDC parameter. The axes for these graphs are identical to those of Fig. 5. However, now we have zoomed in on the data at different points for **a** and **b** to graphically illustrate the RDC. For both a and b, the open circles represent the theoretical relationship between ROI width and 100 * CV. The filled circles represent the actual data. If one starts at the actual data point for the largest ROI (21 × 21), projects a line parallel to the abscissa onto the theoretical line, and then projects down to the abscissa, one can read off the RDC from the abscissa in ROI width units. In the case of graph a, the RDC was 16.6 pixels, whereas for b the RDC was 4.4 pixels.

The percent fluctuation data shown in Fig. 8c indicate that a value of ~0.10% is a typical lower level, and that all stable scanners are below 0.2% on an average basis. This is below or comparable to the expected BOLD changes, which, as noted above, can range up to several percent. However, it is important to note that the percent fluctuation measure is based on a 1 SD unit, and that approximately 30% of the temporal noise distribution lies beyond ±1 SD unit. Also, these measures do not incorporate estimates of physiological noise, which as a general matter will be larger than the thermal noise estimates from a phantom.

As for drift values (absolute value), it appears from Fig. 8d that values of ~0.4% are about as good as one

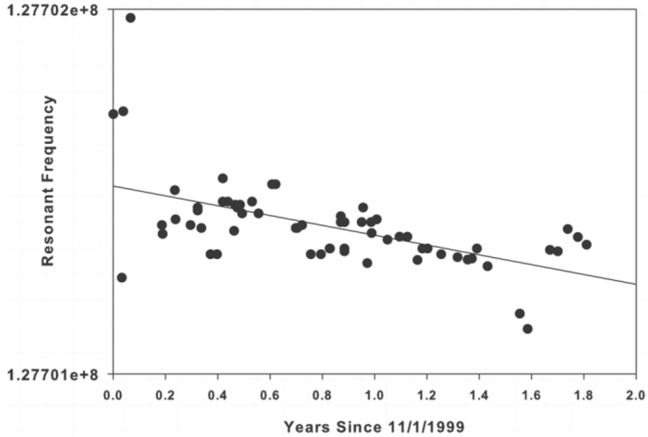

Figure 7. Drift in resonant frequency over time. The decline represents a downward drift in resonant frequency of 135 Hz/year [F(regression) = 31.3, df = 1,66, $P < 0.0001$].

can get, and that stable scanners generally average around 1.0% or less.

Vendor Differences in Drift

We noticed a vendor difference in drift (Fig. 10). The average intensity of an ROI from GE scanners tends to drift down over time, whereas the average intensity tends to drift up on Siemens scanners. This difference

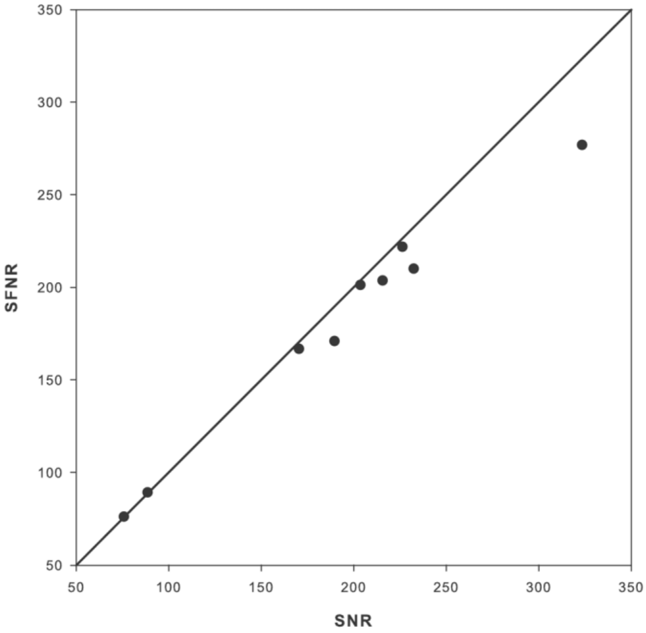

Figure 9. Scatterplot relating the SNR (abscissa) and SFNR (ordinate) to the mean values presented in Fig. 8. The diagonal line is the line of identity.

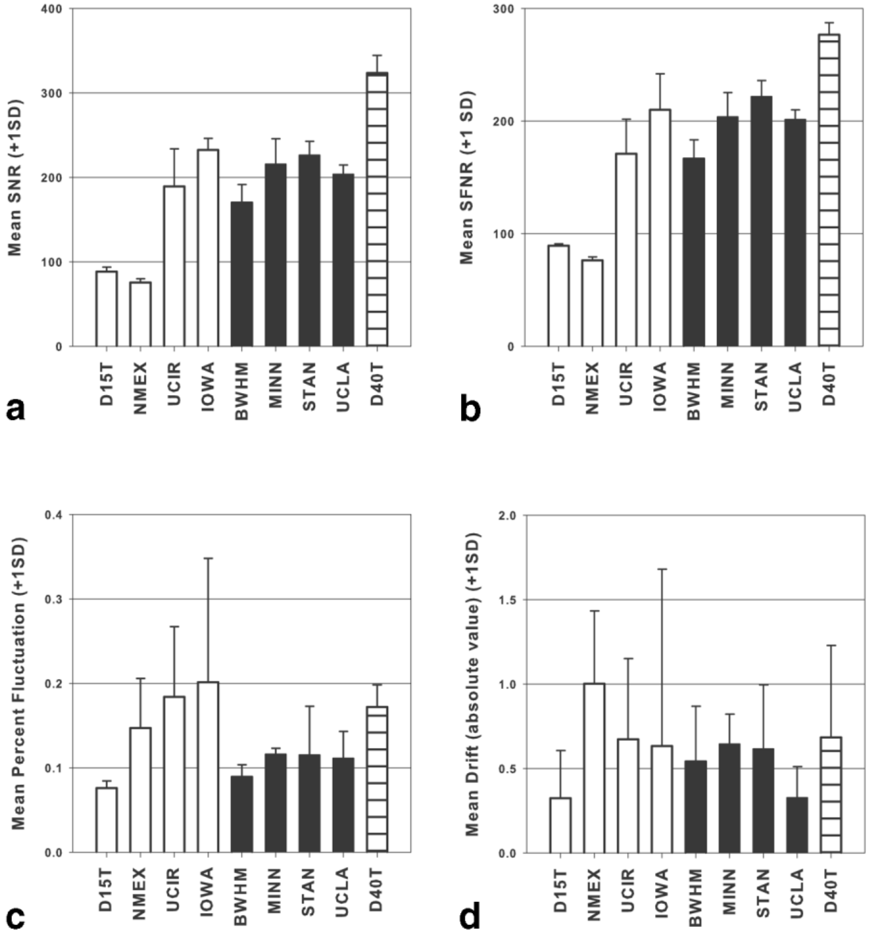

Figure 8. Comparison of nine scanners for four key GSQAP measures. Data are averages (+1 SD) of the last 10 measures from nine scanners that are included in the FIRST-BIRN project. The white bars represent 1.5T scanners, the black bars represent 3T scanners, and the striped bar represents a 4T scanner. **a:** Mean SNR (+1 SD). **b:** Mean SFNR (+1 SD). **c:** Mean percent fluctuation (+1 SD). **d:** Mean drift (absolute value; +1 SD). UCIR and IOWA both employed apodization filters during reconstruction, which increased image smoothness and probably explains their elevated SNR and SFNR.

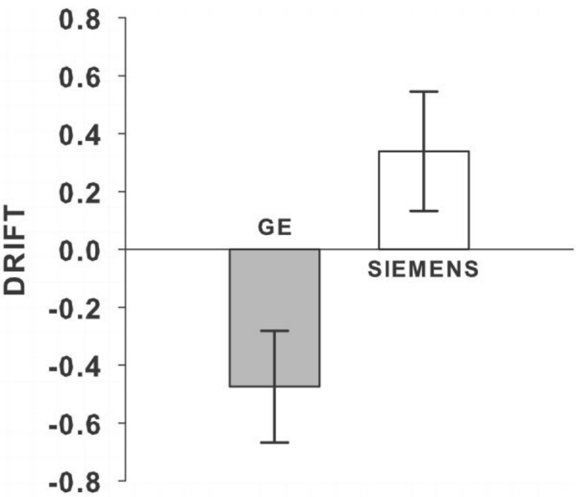

Figure 10. Illustration of vendor differences in the direction of drift. The average drift for all GE scanners is negative, and the average drift for all Siemens scanners is positive. The error bars represent ±1 SD.

was statistically significant (F = 8.3, df = 1,5.2, $P < 0.03$). There was no field-strength effect or a field-strength-by-vendor interaction. To our knowledge, we are the first to report this difference between scanners. We are not certain of the causes of the difference, but because these drifts occur over a period of approximately 10 minutes it is reasonable to suspect some sort of warm-up issue. Perhaps warming up the scanners with several runs of the GSQAP prior to the "official" run would reveal substantially less drift for both vendors. With the downward drift of the GE scanners, instability in the RF-transmit gain could be a likely cause.

Dependence of SNR, SFNR, Percent Fluctuation, and RDC on Image Smoothness

Spatial smoothness can be added to image data at the time of reconstruction due to the application of k-space (a.k.a. "apodization") filters (9). In another report (10) we found that vendor and site effects on image spatial smoothness can have important effects on the activation contrast-to-noise ratio (CNR). Therefore, we were interested in evaluating the dependency of key variables in the GSQAP on the spatial smoothness of the input images (Fig. 11). The data from an "unsmooth" (high-resolution) site (NMEX) were smoothed with Gaussian kernels with full width at half maximum (FWHM) from 1 mm to 15 mm in 1 mm steps. Both SNR and SFNR increased dramatically with increased smoothness, as expected. Percent fluctuation decreased gradually from 0.1 to approximately 0.05 over this smoothness range. Of particular note is the monotonic decline in RDC as the smoothing kernel increases in width. (Drift was stable across smoothness levels, as expected.)

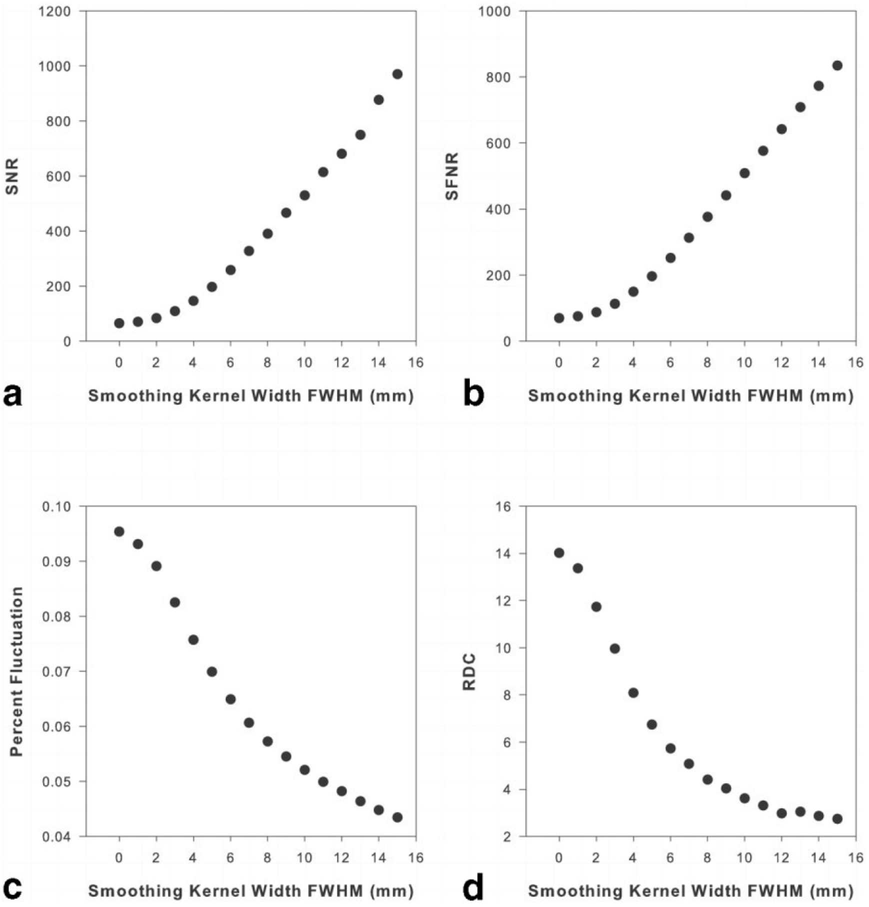

Figure 11. Scatterplots relating the effect of spatially smoothing the phantom time-series on four GSQAP measures. In **a–d** the abscissa is the width, in millimeters, of the Gaussian smoothing kernel applied to the phantom time-series. a: The ordinate plots the SNR. b: The ordinate plots the SFNR. c: The ordinate plots the percent fluctuation. d: The ordinate plots the RDC.

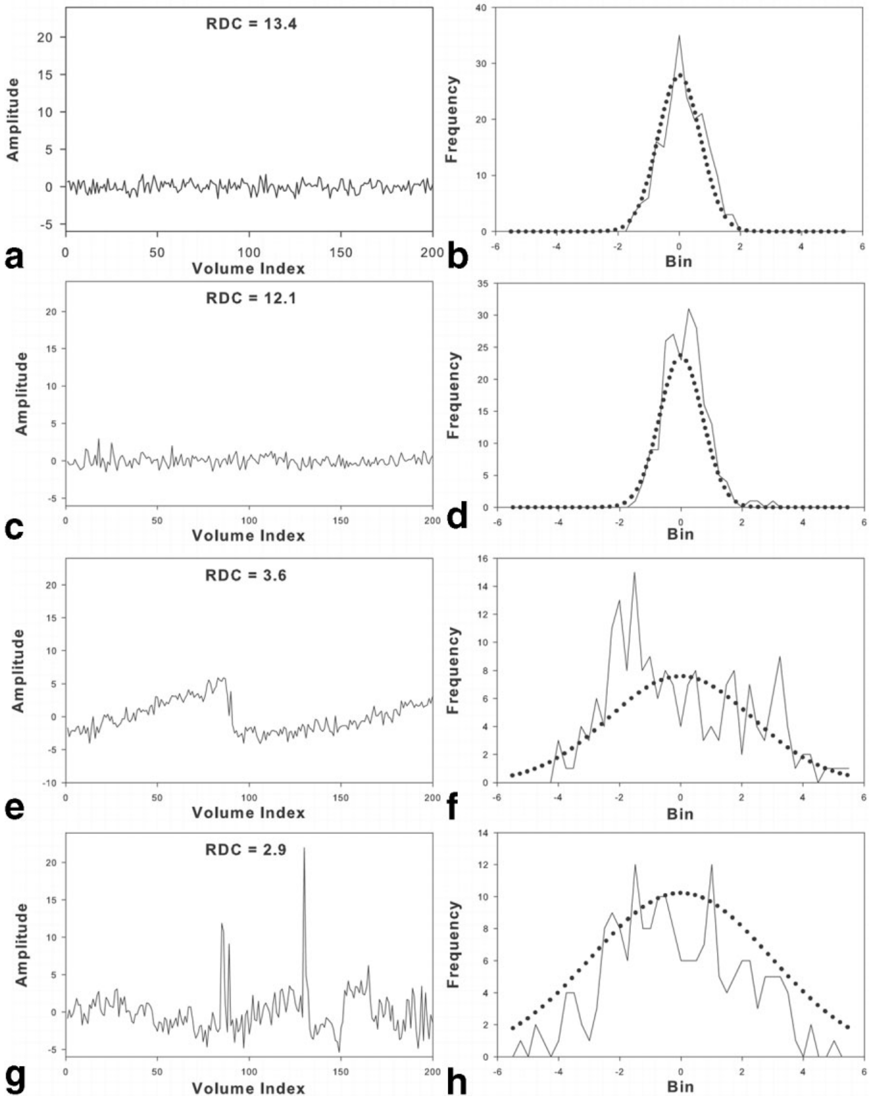

Figure 12. Illustration of the dependence of the RDC measure on the stationarity of the time-series to which it is applied. All data are from one scanner (NMEX). Subplots **a**, **c**, **e**, and **g** are four different time-series after a second-order polynomial trend has been removed. For these subplots the abscissa is the volume number, and the ordinate is the amplitude. Subplots **b**, **d**, **f**, and **h** are frequency histograms of the adjacent time-series (solid line) and a best normal distribution fit (dotted line). See text for additional details.

Interpretation Issues With the Weisskoff Analysis

As noted above, we found a strong relationship between the smoothness of the images and the RDC measure (Fig. 11). What this means is that smoother images (e.g., from a reconstruction with a k-space filter in place) will have a lower RDC than higher-resolution images from another scanner, regardless of noise.

Also, the RDC values appear to be sensitive to irregularities in the time course of the residuals over time (Fig. 12). In Fig. 12a, c, e, and g are four time-courses of the residuals of the 441-voxel ROI. In Fig. 12a and c the time courses appear to be quite stationary and regular, with random fluctuations, and the RDC values are high. In Fig. 12e and g the time courses are irregular and nonstationary, and the RDC values are quite low. The graphs on the right-hand side of Fig. 11 (b, d, f, and h) are histograms of the residuals shown on the left side, with a normal curve shown in dashed lines for comparison. Note that the residuals appear to be more or less normally distributed in b and d (high RDC), but are a poor fit to the normal distribution in f and h (low RDC).

Value of Longitudinal Assessments

Sites in the FIRST-BIRN project have been encouraged to run the GSQAP once a week, and an evaluation of the values collected over the last 80 weeks or so has provided additional insight into the usefulness of this protocol. Bryon Mueller, Ph.D., a physicist at the Center for Magnetic Resonance Research (CMRR) at the University of Minnesota, has been one of the most diligent users of the GSQAP, and some of our best examples come from him and his site (Fig. 13).

One point that this analysis (Fig. 13) makes is that there has been a gradual tightening of the data values for all of the measures from the time the GSQAP was first used (around June 2003) to the present. By having comparative measures of scanner performance at a given site from one week to the next, and across the consortium of sites acquiring QA data with the same protocol, each site has been able to improve and then maintain its stability performance.

For example, the installation of a new transmitter board at the Minnesota site at approximately day 270 clearly increased the SNR and SFNR, and markedly

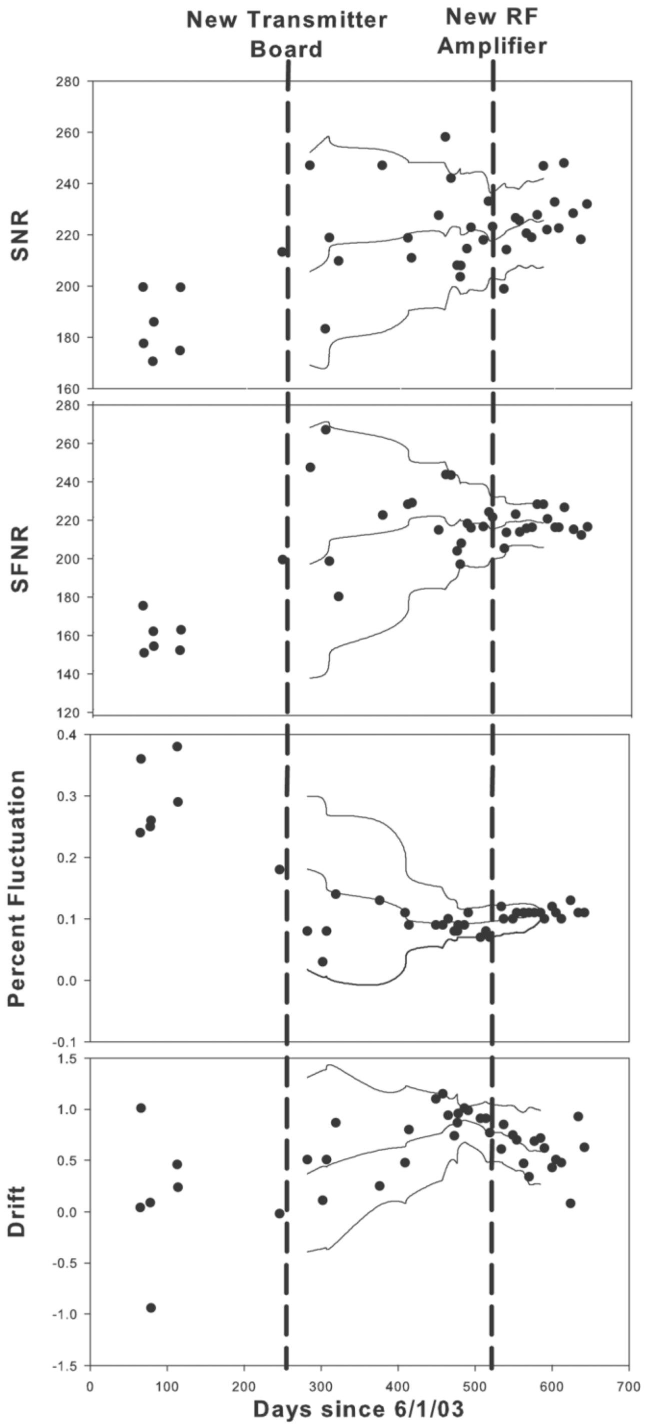

Figure 13. Longitudinal GSQAP assessments from the Minnesota site. The data plotted cover a period of more than 600 days since 6/1/2003. The dark vertical dashed lines that run the length of the figure mark two major hardware changes, noted at the top. The middle line that runs along with the data points on each graph is a centered moving average (length = 15 points), and the lines above and below this middle line are this moving average ±1.5 SD units. The latter lines are included to emphasize the longer-term trends in the data.

reduced percent fluctuation (instability). This is an excellent demonstration of the value of the GSQAP, in that the effect of various hardware changes can be tracked over time. These kinds of objective data can be very useful when one discusses service/performance issues with scanner vendors. The installation of a new RF-transmit amplifier around day 520 was associated with a trend toward higher SNR, and reduced day-to-day variability in SFNR and possibly percent fluctuation. There was also a trend toward reduced drift after this hardware change was made.

A similar analysis from the New Mexico site is illustrated in Fig. 14. The format is the same as in Fig. 13.

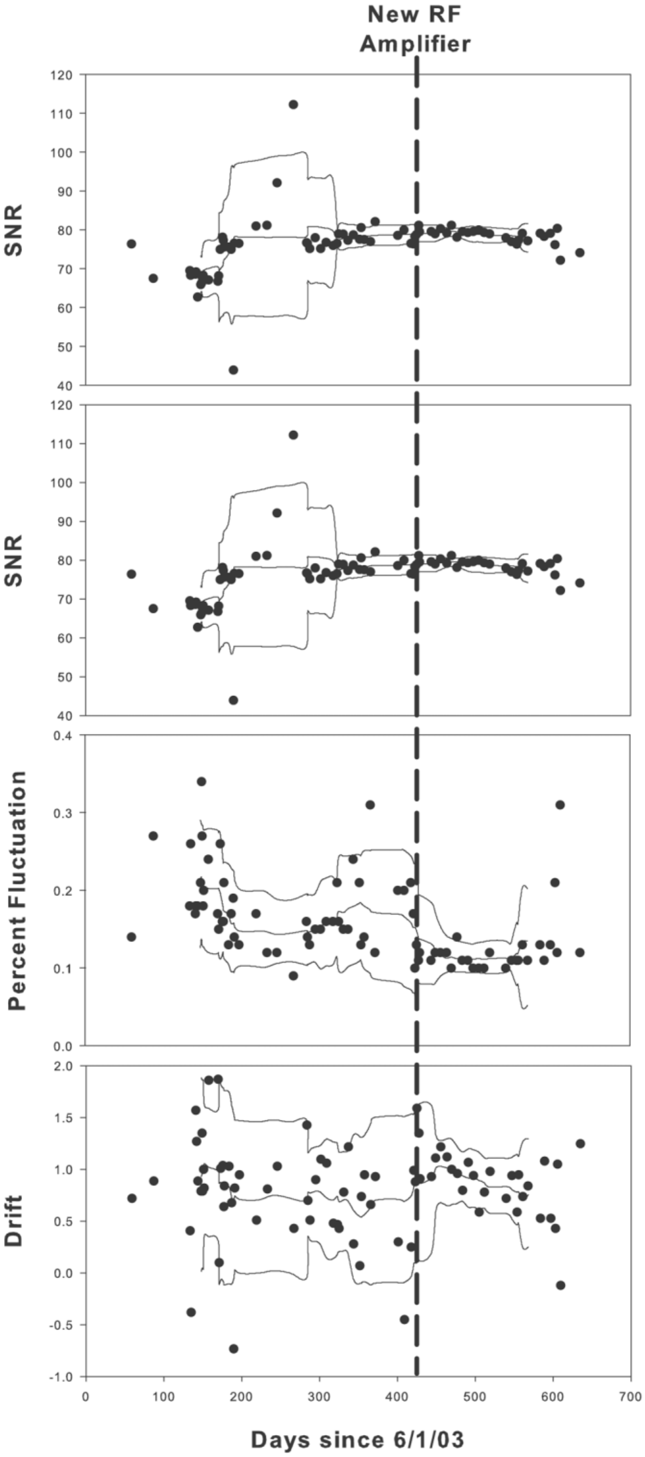

Figure 14. Longitudinal GSQAP assessments from the New Mexico site. See Fig. 13.

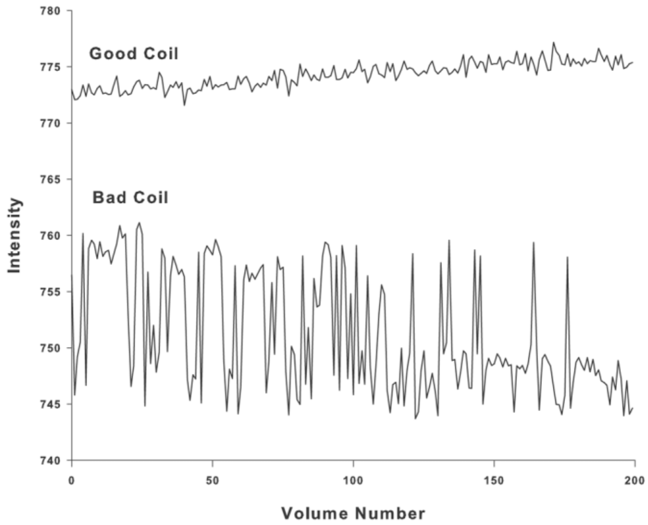

Figure 15. Time-series data from a good coil and a faulty coil.

Once again, there is a general trend for a reduction in variability in all measures over the entire period, but it is less marked than that for the Minnesota site. The SNR and SFNR since about day 300 have been extremely stable on this system. A new RF-transmit amplifier installed on this system on day 423 was followed by a drop in percent fluctuation and a trend toward a decrease in drift. The few recent unusual percent fluctuation values around day 600 were due to dirty head coil contacts (see below).

The Minnesota site has had a repeated problem with head coils, and was able to use the GSQAP to diagnose a bad head coil (Fig. 15). Figure 15 shows this time course with a faulty head coil and after the faulty head coil was replaced. Note the marked nonstationarity (plateaus and spikes) with the faulty coil.

DISCUSSION

Scanner stability is obviously key to successful fMRI research. It is obvious that stability throughout a run (continuous set of volumes) is important because of the low signal (BOLD) that is being measured. Stability at larger scales, i.e, across runs within a study (subject visit); across days, weeks, and months; and between scanners at different sites, can also be quite important depending on the study being conducted. Our results highlight the fact that scanners are always changing, and this has important implications for how one conducts fMRI research. For example, the strategy of collecting controls first and patients later is clearly flawed given the likelihood of scanner performance changes. Long pauses in the conduct of a study can be associated with changes in the performance of the scanner and can add unwanted variance to a study. Site differences in scanner performance will clearly affect the results obtained from the various sites.

Our study provides specific examples regarding the use of various GSQAP measurements. However, it may be instructive to consider more carefully their significance relative to scanner hardware, especially since three of these measures—fluctuation, SNR, and SFNR—appear to be strongly correlated. Indeed, SNR and SFNR usually provide nearly redundant information. SNR is derived from the ratio of the mean of an ROI in the time-series signal image and the variance in an ROI in the DIFF image. SFNR is derived from the ratio of signal and temporal SD maps. In our experience, when they differ by more than a few percent it is because of a low-frequency structure in one of the maps (DIFF, signal, or SD). Such a low-spatial-frequency structure typically results from phase instabilities in either the gradient subsystem (e.g., a noisy gradient amplifier) or any part of the RF subsystem, including the low-power synthesizer/exciter, power amplifier, transmit/receive switch, coil, and receiver. The temporal fluctuation measure and its Fourier spectrum provide a more-detailed look at instability that can be very helpful. Discontinuities in the temporal data can highlight the presence of an intermittent component, frequently in the higher-power parts of the scanner, such as the gradients or RF transmitter, but also in the coil and many other sources. Specific instability frequencies can provide a clue as to the source. For example, a malfunctioning magnet refrigerator component ("cold head") with excess vibration was identified by the presence of its characteristic cycle period as a peak in the spectrum. In other cases, spike noise (very brief transient energy in the detected signal) can be identified in the time-series plots that would not be distinguished as such in the SNR measurements. The Weisskoff measurement of RDC also tends to highlight phase instabilities during the acquisition window.

We have described the experience of a multicenter consortium (FIRST-BIRN) that has used the GSQAP for over two years. As a general matter, our experience has been that the first use of the GSQAP at sites reveals problems in scanner performance. Continued use is associated with a gradual increase in performance and stability of performance.

Regular measurements of SNR, SFNR, percent fluctuation, and drift provide critical feedback regarding scanner performance. As noted above, we have provided realistic guidelines for these measures based on 9 scanners from three vendors, at three field strengths. In the first instance, the data will help others determine whether their scanner is in the normal range. Measurements over time can highlight variability in performance and the effects of hardware and software changes on performance. The Fourier analysis can reveal periodic noise in the time-series, and this information can be useful for diagnosing the source of the noise. Drifts in the RF amplifier gain settings and resonant frequency can also provide valuable feedback about the state of a scanner.

We have also noted a number a ways in which the GSQAP can be improved. A simple improvement that would make the data even more useful would be to require the operator to answer some simple text questions prior to running the analysis, such as: What is the state of the scanner? What kind of problems have been occurring? When was the last significant hardware or software change made? What is the purpose of this run of the GSQAP? What is the current temperature and humidity in the scanner room ? Databasing these text comments into the results would make the data much more useful.

Signal drift over a period of 10 minutes or so suggests sources related to temperature changes or “warming up.” Obviously, we are more concerned with drift in our functional in vivo experiments than with drift in the phantom, and for the phantom-based drift measures to be relevant, the scanner should be in the same state during these assessments as during the fMRI experiments. Typically, acquisition of functional scans will follow other scan types (localizers, anatomical scans, etc.), so it would be reasonable to suggest that the phantom acquisitions should be preceded by similar types of acquisitions. In fact, it might be interesting to run complete protocols with the phantom and compare drift measures from run to run. It is possible that drift could be reduced by a standard warm-up procedure. However, often this type of drift is mitigated in fMRI preprocessing by detrending with nuisance covariates or by high-pass temporal filtering.

We have illustrated two issues that confound the interpretation of the Weisskoff (8) analysis. First, it appears to be quite sensitive to nonstationarities in the time-series. Second, it appears to be quite sensitive to the smoothness of the images. For these reasons, it might be worthwhile to replace the Weisskoff analysis with a straightforward assessment of image smoothness for each time point. One benefit of this is that one obtains another scalar report value: average smoothness. Furthermore, if noise leads to increase spatial correlation in the data, smoothness assessments on each volume will pick this up just as the Weisskoff analysis will.

Stocker et al (11) emphasized the assessment of the shape of the noise distribution from a time-series of phantom volumes. They made the case that the residuals after the time-series data are detrended should be Gaussian, and suggested several metrics to test this. Deviance from a Gaussian distribution on any one slice or any one volume can indicate scanner instability. Future versions of the GSQAP may employ such measurements.

Ghosting is a common problem in fMRI acquisitions, and Simmons et al (12) suggested an interesting method to detect ghosting regions and compute signal/ghost ratios. In their method, the entire background (entire image excluding a mask of the phantom signal) is searched with a 10 × 10 pixel ROI. The position of the ROI with the largest mean value is taken as the “ghost” region. The addition of such a ghost-region measurement and assessment of signal/ghost measures might be worth considering in future development of the GSQAP.

Table 3
Devor et al. (14) Rules for Determining When a System is Beyond Control Limits

Rule	Description of rule
1	A sample is > ± 3SD
2	2 of 3 samples > ± 2SD
3	4 of 5 samples > ± 1 SD
4	8 consecutive samples all above or all below the mean
5	6 or more points forming a linear trend away from the mean
6	8 successive points > ± 1 SD

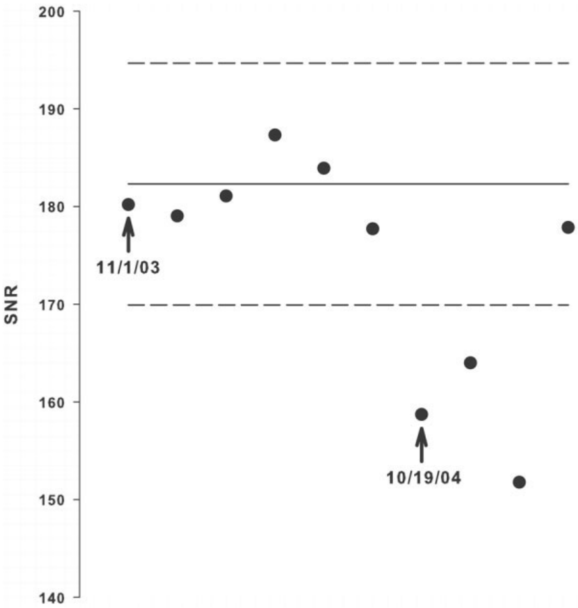

Figure 16. Shewhart chart for data from the BWHM site. The abscissa is the time and the ordinate is the mean SNR. Each black dot represents the mean of five runs of the GSQAP. The solid line represents the mean of the first five dots. The dashed lines represent the upper and lower thresholds, based on ±3 * SEM (based on the first five dots). Note that the system is “out of control” on 10/19/2004. The Shewhart analysis is provided as an option in the SAS statistical software package (SAS Inc., Cary, NC, USA).

Once one is in possession of longitudinal data, the question then arises: When is the system beyond control limits? Simmons et al (12) suggested the use of Shewart charting. Shewhart (13) developed the concepts of statistical process control and provided a set of statistical methods for assessing system stability in the context of manufacturing quality control. Devor et al (14) continued this line of thought. The basic idea is to take samples over time (e.g., five contiguous runs or so per sample) and plot the means. The next step is to compute the mean of a set (e.g., a group of five) means and the standard error of the means (SEM). Confidence limits, based on ±3 SEM, are placed around the grand mean. The occurrence of any sample mean outside of these confidence limits indicates a system that is beyond the control limits (i.e., out of specification). Shewhart (13) developed similar tests for the variability (range and SD) and distributional shapes (skewness and kurtosis) of the samples as well. Other specific rules for determining when a system is beyond control limits were proposed by Devor et al (14) and are given in Table 3. We applied this approach to SNR data from the BWHM site (Fig. 16). According to this method, their system state changed on 10/19/2004. It turns out that about that time, the BWHM site began to use a “shell loader” designed to increase the coil load of a phantom. Since the recommended phantom for the GSQAP is already loaded to match the coil load of a typical head, there was a drop in SNR during the measurement. The Shewhart (13) approach would have notified these researchers soon after the change. As it happens, they were not made aware of the effect of this change until the present manuscript was being prepared.

For MRI centers considering the initiation of a new QA program, we recommend at least a weekly assessment with the GSQAP. We suggest that users compare the results with those presented in Fig. 8 to get a sense of how they compare to these nine scanners. Any deviation from acceptable performance, as outlined above, should be examined and its cause determined. We have provided some guidance with regard to hardware and reconstruction issues that can affect the measurements, but obviously those involved must work with their local service personnel to track down the issues, since there are so many possible causes of malfunction with this complex hardware. We encourage local users to familiarize the local service personnel with the GSQAP and its benefits as outlined herein. We recommend that sites keep time-stamped logs of GSQAP results and plots, and develop a system to regularly check the stability of the data, as suggested above. Also, it is very important to check the GSQAP results carefully before and after any hardware or software change. Finally, we recommend that fMRI investigators consider stability issues carefully when conducting studies over long periods, given the likelihood of some drift in performance over time. These steps should enhance the probability that the fMRI data collected are of the highest quality.

In conclusion, the GSQAP has become an accepted standard of excellence for scanner performance in the FIRST-BIRN program because experiences such as the examples given above have indicated the value of attaining and adhering to a set of minimum stability requirements. It can be used for evaluation and acceptance of a new scanner, since guidelines for measurements are now available. In fact, at least one site within the consortium used aspects of the specifications for stability as conditions for acceptance of a new scanner from their vendor. The availability of an ongoing database of GSQAP results will be of great value for service personnel in that it will provide objective data and reasonable performance targets. Such a database will also allow for the monitoring of software and hardware upgrades. Finally, it can be quite helpful in the planning of multicenter studies, since differences in stability and performance can have marked effects on results from various sites.

APPENDIX

Stanford Agar Phantom Recipe

G.H. Glover 2/20/2003

The phantom is constructed using 17.5-cm-diameter spherical container from Dielectric. The basic relaxation agent recipe is from Schneiders (15) and is designed to provide T1 and T2 comparable to those in gray matter. A small amount of NaCl is added to increase the conductivity to mimic the RF load of a head.

1. Make 21.8 mM $NiCl_2$ mixture:
   2.82 g of nickel chloride per 1 liter H_2O
2. Make agar mixture:
   3600 mL H_2O
   400 mL 21.8 mM $NiCl_2$
   120 g of agar
   20 g of NaCl (0.5%)
   1 g of sodium azide (toxic, used to retard the growth of evil green things)
3. Boil mixture slowly.
   Use two 2-liter beakers, divide the mixture into halves.
   Boil each liter of mixture as follows:
   Heat the beakers one at a time for three to five minutes at high power in a kitchen-sized microwave (ours is listed as 750 watt; times may (or may not) scale with power). Be careful that it doesn't boil over, and titrate the time accordingly, reducing the cycle time as the temperature rises. To avoid burning the gel, do not use a heating plate to heat the beakers.
   Remove beaker and stir. Put other beaker in oven, swapping back and forth.
   Repeat heat/stir cycles until all of the agar is dissolved and the liquid is light brown but clear. This will take about one to two hours of heating.
   Interleave the beakers throughout the heat/stir cycles unless the microwave is large enough for both at once.
   While the mixture is boiling hot, pour it into the phantom using a funnel.
4. After the final pour, purge all air bubbles using a 50-cc syringe that is filled with liquid and connected to tubing inserted into the bottom of the phantom. Fill the sphere until liquid runs out the hole. Doing this will waste a lot of mixture, so be prepared to catch the spillover. Plug the hole with a nylon screw and o-ring gasket while it is hot.
   Note that the starting volume of the mixture is greater than $4/3\pi r^3$ = 2800 mL. Reduction occurs by boiling of water, spillover, or by magic; however, there will be at least 500 mL left over.
   Courtesy Anne Marie Sawyer-Glover, Lucas Center, Stanford Radiological Sciences Laboratory.

ACKNOWLEDGMENTS

The authors used data from the FBIRN consortium for some of the results presented here. The entire FBIRN consortium provided invaluable assistance. We particularly thank the PI, Dr. Steven Potkin, for his support of this research. We thank Anne Sawyer-Glover for developing the agar phantom recipe and acquiring data for the Stanford site, and Bryon Mueller for discussions about his site. Chuck Gasparovic and John Rasure read earlier versions of the manuscript and provided key feedback.

REFERENCES

1. Ogawa S, Tank DW, Menon R, et al. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proc Natl Acad Sci USA 1992;89: 5951-5955.

2. Matthews PM. An introduction to functional magnetic resonance imaging of the brain. In: Jezzard PM, Matthews PM, Smith SM, editors. Functional MRI—an introduction to methods. Oxford: Oxford University Press; 2001. p 3-34.
3. Ihalainen T, Sipila O, Savolainen S. MRI quality control: six imagers studied using eleven unified image quality parameters. Eur Radiol 2004;14:1859-1865.
4. Price RR, Axel L, Morgan T, et al. Quality assurance methods and phantoms for magnetic resonance imaging: report of AAPM nuclear magnetic resonance Task group no. 1. Med Phys 1990;17:287-295.
5. Thulborn. Quality assurance in clinical and research echo planar functional MRI. In: Moonen C, Bandettini P, editors. Functional MRI. New York: Springer; 2000. p 337-346.
6. National Electrical Manufacturers Association (NEMA). Determination of signal-to-noise ratio (SNR) in diagnostic magnetic resonance images. Nema: Rosslyn, VA, 1988.
7. Meigen T, Bach M. On the statistical significance of electrophysiological steady-state responses. Doc Ophthalmol 1999;98:207-232.
8. Weisskoff RM. Simple measurement of scanner stability for functional NMR imaging of activation in the brain. Magn Reson Med 1996;36:643-645.
9. Lowe MJ, Sorenson JA. Spatially filtering functional magnetic resonance imaging data. Magn Reson Med 1997;37:723-729.
10. Friedman L, Glover GH, Kvenz DE, Magnotta V and the FIRST BIRN Consortium. Reducing scanner-to-scanner variability of activation in a multi-center study. Role of smoothness equalization. Neuroimage (in press).
11. Stocker T, Schneider F, Klein M, et al. Automated quality assurance routines for fMRI data applied to a multicenter study. Hum Brain Mapp 2005;25:237-246.
12. Simmons A, Moore E, Williams SC. Quality control for functional magnetic resonance imaging using automated data analysis and Shewhart charting. Magn Reson Med 1999;41:1274-1278.
13. Shewhart WA. Economic control of quality of manufactured product. New York: D. Van Nostrand Co; 1931.
14. Devor RE, Chang T, Sutherland J. Statistical quality design and control: contemporary concepts and methods. New York: Prentiss-Hall; 1992. 809 p.
15. Schneiders NJ. Solutions of two paramagnetic ions for use in nuclear magnetic resonance phantoms. Med Phys 1988;15:12-16.