Evaluation of inter-site bias and variance in diffusion-weighted MRI

Allison E. Hainline[a], **Vishwesh Nath**[b], **Prasanna Parvathaneni**[a], **Justin Blaber**, **Baxter Rogers**[c], **Allen Newton**[c,d], **Jeffrey Luci**[e,f,g], **Heidi Edmonson**[h], **Hakmook Kang**[a,k], and **Bennett A. Landman**[b,i,j,k]

[a]Biostatistics, Vanderbilt University, Nashville, TN 37212

[b]Computer Science, Vanderbilt University, Nashville, TN 37212

[c]Institute of Imaging Science (VUIIS), Vanderbilt University, Nashville, TN 37212

[d]Radiology and Radiological Sciences, Vanderbilt University, Nashville, TN 37212

[e]Neuroscience, The University of Texas at Austin, Austin, TX 78712

[f]Biomedical Engineering, The University of Texas at Austin, Austin, TX 78712

[g]Imaging Research Center, The University of Texas at Austin, Austin, TX, 78712

[h]Radiology, Mayo Clinic, Rochester, MN, 55905

[i]Electrical Engineering, Vanderbilt University, Nashville, TN 37212

[j]Vanderbilt University Institute of Imaging Science, Vanderbilt University, Nashville, TN 37212

[k]Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37212

Abstract

An understanding of the bias and variance of diffusion weighted magnetic resonance imaging (DW-MRI) acquisitions across scanners, study sites, or over time is essential for the incorporation of multiple data sources into a single clinical study. Studies that combine samples from various sites may be introducing confounding due to site-specific artifacts and patterns. Differences in bias and variance across sites may render the scans incomparable, and, without correction, any inferences obtained from these data are misleading. We present an analysis of the bias and variance of scans of the same subjects across different sites and evaluate their impact on statistical analyses. In previous work, we presented a simulation extrapolation (SIMEX) technique for bias estimation as well as a wild bootstrap technique for variance estimation in metrics obtained from a Q-ball imaging (QBI) reconstruction of empirical high angular resolution diffusion imaging (HARDI) data. We now apply those techniques to data acquired from 5 healthy volunteers on 3 independent scanners under closely matched acquisition protocols. The bias and variance of GFA measurements were estimated on a voxel-wise basis for each scan and compared across study sites to identify site-specific differences. Further, we provide model recommendations that can be used to determine the extent of the impact of bias and variance as well as aspects of the analysis to account for these differences. We include a decision tree to help researchers determine if model adjustments are necessary based on the bias and variance results.

1. INTRODUCTION

Given pragmatic considerations of study design and magnetic resonance imaging (MRI) data acquisition, many clinical studies combine data from several different sources in order to increase the sample size and improve the power [1, 2]. Traditional techniques for evaluating contrasts and testing differences across time assume that the bias and variance are constant across all acquisitions. However, there is no universal technique for evaluating sources of bias and variance in MRI on individual subjects. Violation of statistical assumptions has the potential to invalidate inferences. Thus, the addition of new data has the potential to *decrease* the statistical power as a result of the introduction of bias. Significant amounts of bias and variance can result even within a single site due to patient factors, hardware differences, and signal processing/software.

Here, we focus on the context of high angular resolution diffusion imaging (HARDI), with a specific focus on Q-ball imaging (QBI). To illustrate the problem, Figure 1 presents the variation that can be observed within a single subject scanned across 5 separate scans (3 independent scanners and 2 re-scans). Each of these scans was taken under comparable acquisition parameters and should reveal the same brain structure; however, the figure shows variation in B_0, DWI, and vector-mapped images. These differences are clearly visually appreciated from the images themselves, but the extent of the variation's impact on HARDI analysis is difficult to quantify visually.

In diffusion tensor imaging (DTI), bias and variance haven been assessed for single subjects with simulation extrapolation (SIMEX) and Monte Carlo methods, respectively [3]. Recently, these methods have been adapted to HARDI [4], but have not been evaluated on multi-site traveling data. This manuscript presents an analysis of 3 sites using harmonized HARDI acquisition protocols. The focus of this work is to consider tools for estimating the bias and variance, as well as to present a decision tree to guide the researcher as to how such data could be used. The process described herein can be used for quality assurance within a single site to ensure optimal statistical power and results.

2. METHODS

2.1 Data acquisition

Subjects were imaged at 3 independent study sites. Five subjects were imaged at Site A on a 3.0T system using a full body transmit coil with a 32 channel head only receive coil. Non-diffusion weighted imaging sequences consisted of 3D T1 weighted MPRAGE, resting state fMRI, and B0 mapping. Diffusion weighted imaging sequences consisted of a 96 direction DTI (b=1000, 1500, 2000, 2500; SENSE = 2.5; partial Fourier factor = .77; vox. dim. = 1.9×1.9mm^2; FOV = 112×112; # of sl. = 48, slice thickness = 2.5mm), as well as regularly interspersed acquisitions of a three direction DWI acquisition acquired with reversed phase

encoding gradients, and finally a vendor standard 30 direction DTI (b=1000) acquisition. Scans from this site were resampled to $2.5\times2.5\text{mm}^2$ for comparison with the other two study sites. 4 subjects were re-scanned with the same protocol.

Five subjects were imaged at Site B on a 3.0T system using a full body transmit coil with a 32 channel head only receive coil. Non-diffusion weighted imaging sequences consisted of 3D T1 weighted MPRAGE, resting state fMRI, and B0 mapping. Diffusion weighted imaging sequences consisted of a 96 direction DTI (b=1000, 1500, 2000, 2500; SENSE = 2.5; partial Fourier factor = .77; vox. dim. = $2.5\times2.5\text{mm}^2$; FOV = 96×96; # of sl. = 48, slice thickness = 2.5mm), as well as regularly interspersed acquisitions of a three direction DWI acquisition acquired with reversed phase encoding gradients, and finally a vendor standard 30 direction DTI (b=1000) acquisition. 4 subjects were re-scanned with the same protocol.

Four subjects were imaged at Site C on a 3.0T system using a full body transmit coil with a 32 channel head only receive coil. Non-diffusion weighted imaging sequences consisted of 3D T1 weighted MPRAGE, resting state fMRI, and B0 mapping. Diffusion weighted imaging sequences consisted of a 96 direction DTI (b=1000, 1500, 2000, 2465; GRAPPA = 2; vox. dim. = $2.5\times2.5\text{mm}^2$; FOV = 96×96; # of sl. = 50, slice thickness = 2.5mm), as well as regularly interspersed acquisitions of a three direction DWI acquisition acquired with reversed phase encoding gradients, and finally a vendor standard 30 direction DTI (b=1000) acquisition.

2.2 Model fitting

We fit all data with a Q-ball imaging with a model order 6 reconstruction of the orientation distribution function (ODF) of the HARDI data acquisitions. As detailed in [5] and [6], we use a regularized spherical harmonic reconstruction of the ODF and calculate generalized fractional anisotropy (GFA),

$$GFA = \frac{std(\psi)}{rms(\psi)} = \sqrt{\frac{n \sum_{i=1}^{n} (\psi_i - \bar{\psi})^2}{(n-1) \sum_{i=1}^{n} \psi_i^2}} \quad (1)$$

where ψ is the ODF vector, and $\bar{\psi}$ is its mean [7].

Note that we have chosen to use GFA as the metric of interest for this analysis. However, the theory behind these methods allows for calculations based on any scalar metric, with the only requirements being that it is continuous and monotonic with respect to the addition of noise [8].

2.3 GFA bias estimation

We used the SIMEX approach to estimate the bias of GFA. This approach was adapted from modern statistical methods [8] and has been described in [3] for use in DTI, and in progress

for HARDI [4]. When measurement error is present, the true data is unable to be observed, i.e., we instead observe

$$X_{obs} = X_{truth} + \sigma_E R \quad (2)$$

where X_{truth} is the noiseless truth data, σ_E is the standard deviation of the error distribution, and R is a random sample from a Rician distribution with mean zero and standard deviation of 1. In short, the SIMEX procedure relies on the behavior of the metric of interest as a function of the addition of random noise. As noise is added to the observed data in increasing amounts, a trend is observed in the metric of interest calculated at each level of noise. We can use this trend to extrapolate backward to the case with no measurement error (or noise) and obtain a function of X_{truth} [8].

SIMEX does not require the fitting of parametric measurement error models in order to estimate the bias [9]. The only requirements for the application of SIMEX are that the measurement error variance can be estimated and that the metric of interest is smooth and monotonic as a function of noise.

2.4 GFA variance estimation

The wild bootstrap was used to estimate the variance of GFA. The wild bootstrap, as detailed in [10], is a method for estimating the variance of an MRI-derived metric, without requiring the use of several repeated data acquisitions. We use the wild bootstrap rather than a traditional bootstrap resampling with replacement due to the heteroscedasticity of the errors in a diffusion model[11, 12].

2.5 Analysis

For this analysis, we have data acquired from two shells, b=1000 and b=2500. These shells will be analyzed separately and results are compared. The SIMEX and bootstrap procedures are performed on a voxel-by-voxel basis, where each voxel is evaluated independently. All calculations were performed in Matlab version R2016a [13] and the Camino Diffusion MRI toolkit [14].

The first step in the analysis is the estimation of both the bias and variance (as detailed in sections 2.3 and 2.4) for each voxel within each data acquisition. For evaluation, all data was registered to the ICBM 2009a Nonlinear Symmetric template [15, 16]. We chose to do our analysis on the ROI level; thus, we used manually delineated five white matter ROIs: centrum semiovale, splenium of the corpus callosum, internal capsule, putamen, and globus pallidus. The average bias and variance values were taken within ROIs, resulting in a value for each subject, scanner, b-shell, and ROI combination.

When the average values for the bias and standard deviation of GFA are calculated for each ROI, the decision tree found in Figure 2 can be used to guide the model selection process. First, we create quantile-quantile (Q-Q) plots and histograms for both the average bias and average standard deviation of GFA values within each ROI. Q-Q plots compare the quantiles

of the observed distribution to that of a Gaussian distribution. Ideally, the points on the Q-Q plots fall directly on the line, though slight deviations are often not cause for alarm. Often, deviations that occur in the tails of the distribution may be due to small sample sizes and are expected to stabilize with larger samples. Note that formal tests of normality are available; however, these tests can be misleading [17]. Thus, we recommend visually checking for normality via the Q-Q plots and histograms. Bimodality or other signs of asymmetry will be evident in both types of plots. If the plots reveal non-normality, the recommendation is to closely examine the raw data for artifacts and correct them, if found. If no artifacts are found, non-parametric methods that do not assume normality should be used.

Next, boxplots are made for the mean values of the bias and standard deviation of GFA for each ROI with the points overlaid. Each point represents the average value within the ROI for either the estimated bias of GFA or the estimated standard deviation of GFA. The boxplot allows for the visual identification of outliers and gives a clear picture of the differences across sites and subjects. This boxplot is the main tool for determining what modeling strategy is optimal for the data distributions.

3. RESULTS

This analysis consisted of 5 subjects who were scanned at up to 3 independent scan sites with re-scans. Our first step was to create bias and standard deviation maps, shown in Figure 3 for subject 01. These maps help reveal the spatial distributions of both the bias and the variance across the brain structures. In particular, we see higher bias and standard deviations in the gray matter in comparison to white matter.

Following the decision tree in Figure 2, we made Q-Q plots and histograms of the ROI-averaged data (Figure 4) and found that our data were reasonably close to a Gaussian distribution. The plots reveal slight departures from normality due to heavier tails, though the distributions maintain a level of symmetry. In particular, we see a larger negative bias in the internal capsule of a single scan as well as larger standard deviations for the splenium of the corpus callosum for several scans. A look into the raw data revealed that faulty segmentation was at fault and should be fixed for future analyses. Non-parametric methods may be useful for analyses with these slight departures from normality coupled with small sample sizes, though they are not required.

We then created a scatterplot with the bias across the x-axis and the variance across the y-axis to assess the spread of the data and identify outliers (Figure 5). This scatterplot shows deviations of subject 00 at Scanner A from the other data points in terms of average bias for b=1000 as well as several subjects and sites for b=2500. In addition, two scans from Site B show larger than expected standard deviations. These outlying data points should be examined thoroughly before continuing the analysis.

Finally, we made boxplots for both bias and standard deviation of GFA. Figure 6 shows these boxplots for both b=1000 and b=2500. Each data point represents an averaged value across a single ROI for one subject at one site. This figure helps reveal patterns in bias and variance that may be due to either subject-specific variation or differences in acquisition

protocols. According to our proposed rule of thumb, subject 00 at site A has larger than average bias in the internal capsule (b=1000). For b=2500, several scans appear to be outliers in terms of bias in the internal capsule. These results prompt the inclusion of a random effect for subject.

4. DISCUSSION

This manuscript demonstrates the importance of a review of scan quality when combining data from several study sites. With our method of scanning the same subject at multiple scan sites, we are able to estimate biases and variances that can be attributed only to differences in scan conditions, rather than subject-to-subject heterogeneity. The methods described in this paper may be used to analyze the data quality of any study where heterogeneity between sites, scanners, or subjects is a concern. The decision tree in Figure 2 may be used to help guide the model selection process for analyses that combine data from different sources or where there is thought to be differences in the quality between different acquisitions.

Acknowledgments

References

1. Di Martino A, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry. 2014; 19(6):659–67. [PubMed: 23774715]

2. Jack CR Jr, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging. 2008; 27(4):685–91. [PubMed: 18302232]

3. Lauzon CB, et al. Assessment of bias in experimentally measured diffusion tensor imaging parameters using SIMEX. Magn Reson Med. 2013; 69(3):891–902. [PubMed: 22611000]

4. Hainline AE, Nath V, Parvathaneni P, Blaber JA, Schilling KG, Anderson AW, Kang H, Landman BA. Empirical Single Sample Quantification of Bias and Variance in Q-ball. Submitted.

5. Hess CP, et al. Q-ball reconstruction of multimodal fiber orientations using the spherical harmonic basis. Magn Reson Med. 2006; 56(1):104–17. [PubMed: 16755539]

6. Descoteaux M, et al. Regularized, fast, and robust analytical Q-ball imaging. Magn Reson Med. 2007; 58(3):497–510. [PubMed: 17763358]

7. Tuch DS. Q-ball imaging. Magn Reson Med. 2004; 52(6):1358–72. [PubMed: 15562495]

8. Cook JR, Stefanski LA. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statistical Association. 1994; 89(428):1314–1328.

9. Carroll RJ, et al. Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. Journal of the American Statistical Association. 1996; 91(433):242–250.

10. Jones DK. Tractography gone wild: probabilistic fibre tracking using the wild bootstrap with diffusion tensor MRI. IEEE Trans Med Imaging. 2008; 27(9):1268–74. [PubMed: 18779066]

11. Basser PJ, Mattiello J, LeBihan D. Estimation of the effective self-diffusion tensor from the NMR spin echo. J Magn Reson B. 1994; 103(3):247–54. [PubMed: 8019776]

12. Whitcher B, et al. Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging. Hum Brain Mapp. 2008; 29(3):346–62. [PubMed: 17455199]

13. Matlab. MATLAB Release 2016a. The MathWorks, Inc; Natick, MA, United States:

14. Cook, PA., YB, Nedjati-Gilani, S., Seunarine, KK., Hall, MG., Parker, GJ., Alexander, DC. Camino: Open-Source Diffusion-MRI Reconstruction and Processing. 14th Scientific Meeting of the International Society for Magnetic Resonance in Medicine; 2006; Seattle, WA.

15. Fonov VS, ACE, Botterton K, Almli CR, McKinstry RC, Collins DL. BDCG. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage. 2011; 54(1)

16. Fonov VS, ACE, Almli CR, Collins DL. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage. 2009; 47(Supplement 1)

17. Mason, DRaS, JH. A Modified Kolmogorov-Smirnov Test Sensitive to Tail Alternatives. Ann Statist. 1983; 11(3):933–046.
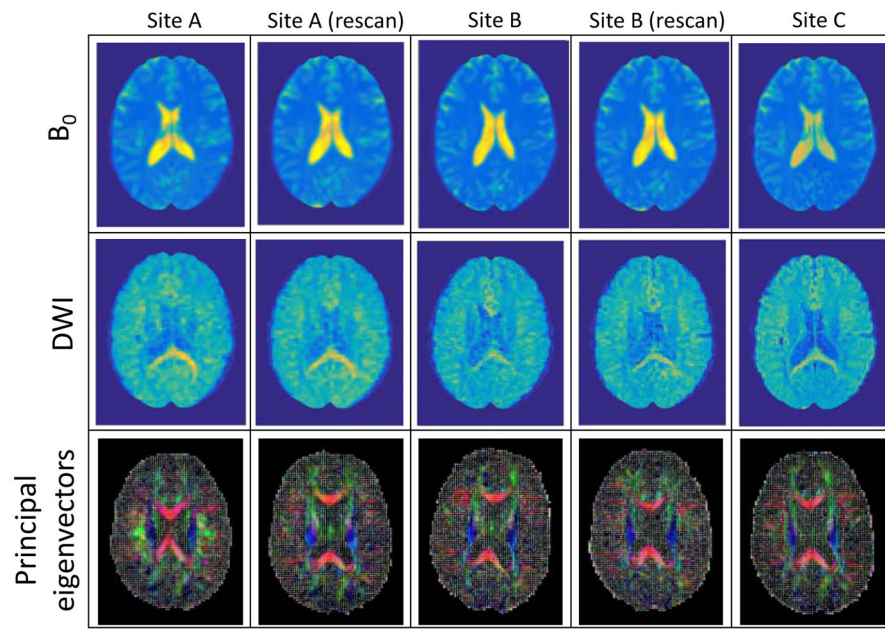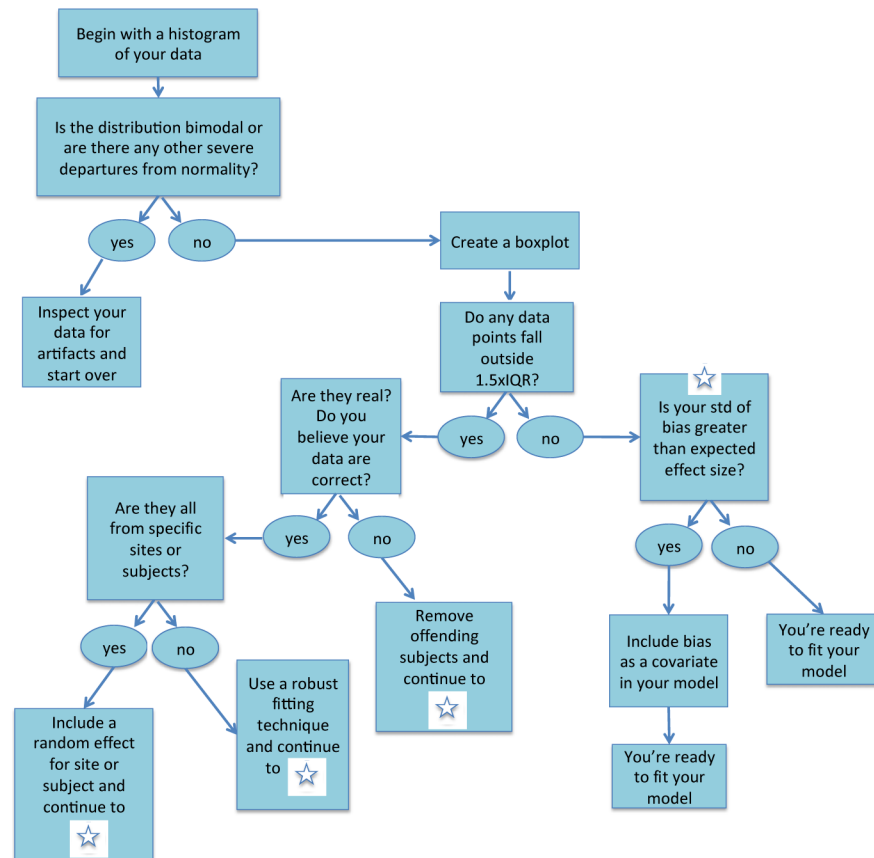
Figure 1.
Illustration of the variation that may be seen across scan sites within a single subject with each scan processed in the acquired space. B_0, DWI, and the principal eigenvectors for the same mid-axial slice across different scans within the same subject. The color maps are constant across all sites for each.

Figure 2.
Recommended decision-making process for model selection. The procedure begins with a simple histogram or density plot to identify any severe distributional issues that would interfere with inference. For the purpose of this decision tree, the "data" refers to the averaged values within each ROI for either the estimated bias of the metric of interest or the estimated standard deviation of the metric of interest. Once the distribution has been checked, a boxplot is recommended to get an idea of what outliers exist and where they came from. Any patterns in outliers that can be attributed to site- or subject-specific artifacts should be accounted for with a random effect. Finally, one should look at the standard deviation of the metric as well as the magnitude of the bias. If either of these is larger than the expected effect size, this should be accounted for in the modeling through the inclusion of a covariate.
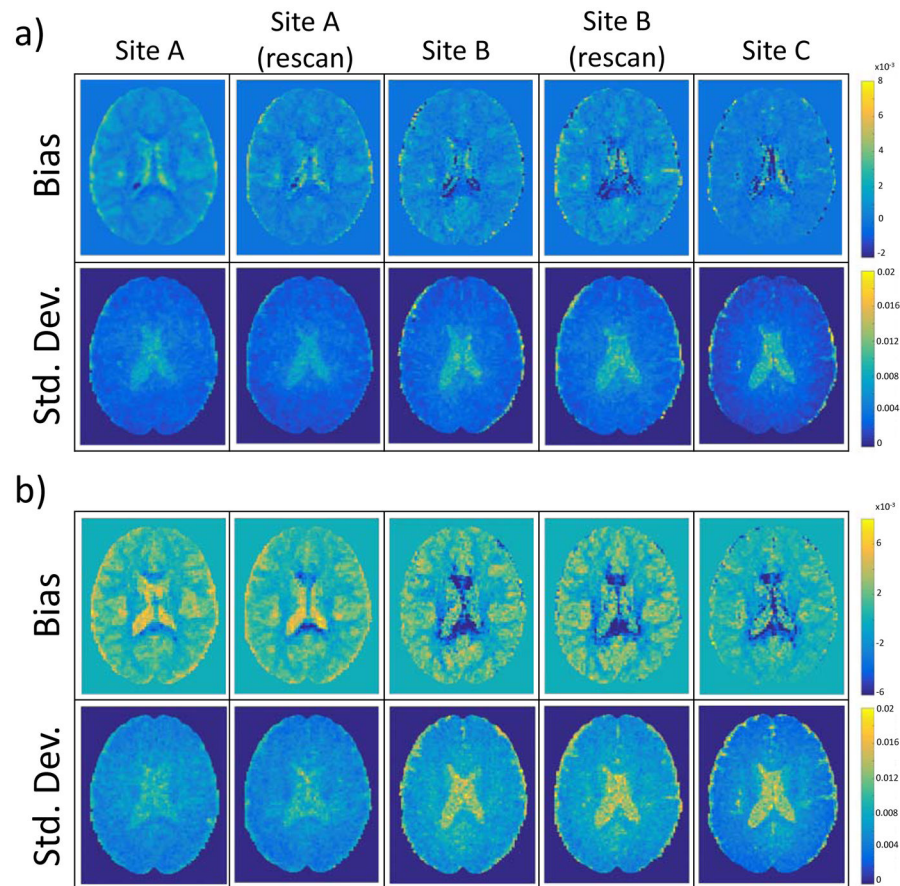
Figure 3.
Spatial maps for the bias of GFA and the standard deviation of GFA for subject 01 are shown across all sites for (a) b=1000 and (b) b=2500. With these maps, the spatially dependent nature of the bias and variance estimates is evident. We can also see the variation in both estimates across the different scans.
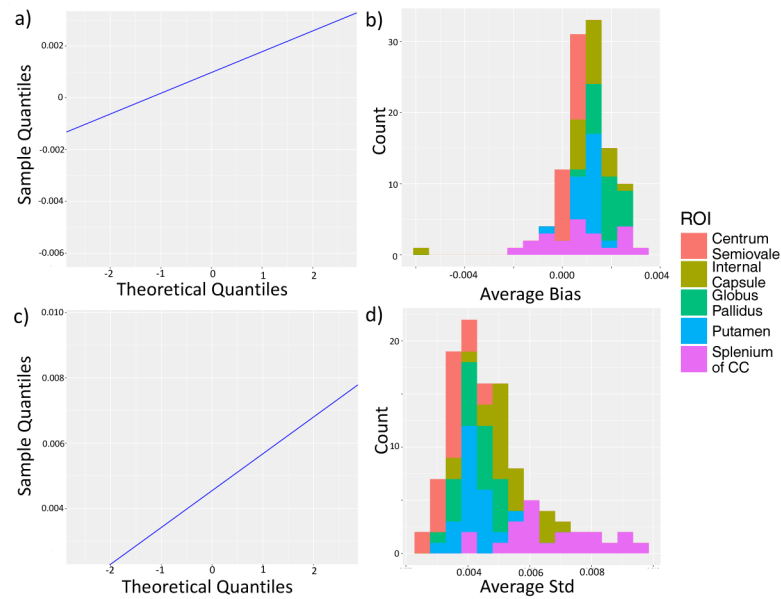
Figure 4.
Checking for normality. For b=1000, Q-Q plot for bias of GFA (a) and histogram for bias of GFA (b) as well as the Q-Q plot for the standard deviation of GFA (c) and the histogram of the standard deviation of GFA (d) reveal distributions with minor deviations from normal distributions, though not different enough to warrant any concern for lack of normality. The bias histogram reveals an outlier in the internal capsule, while the splenium of the corpus callosum appears to have higher than expected standard deviation. These deviations from expectation prompt a deeper look into the images to determine if there are issues in the data, such as artifacts or faulty segmentation.
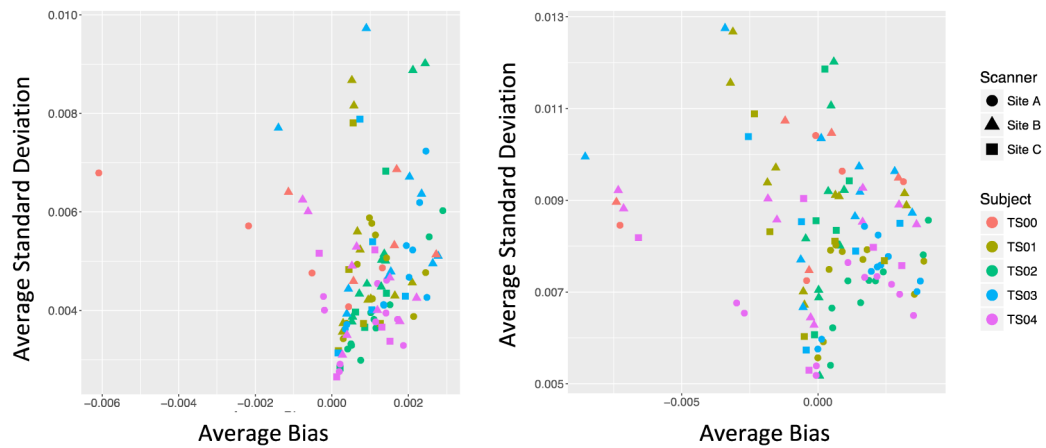
Figure 5.
Scatterplots of average bias and average standard deviation for b=1000 (left) and b=2500 (right). Each data point is the average value within an ROI for each subject/scanner pair. In this plot it is clear that one ROI from subject 00 at Site A has a larger negative bias than the rest of the scans at b=1000. We also find several ROIs from a variety of subjects and sites have larger negative biases than the rest of the scans at b=2500, as well as two scans from Site B which have larger than expected standard deviations. These data should be examined to ensure that these values are correct and that the quality is compatible with the remaining scans in this set.
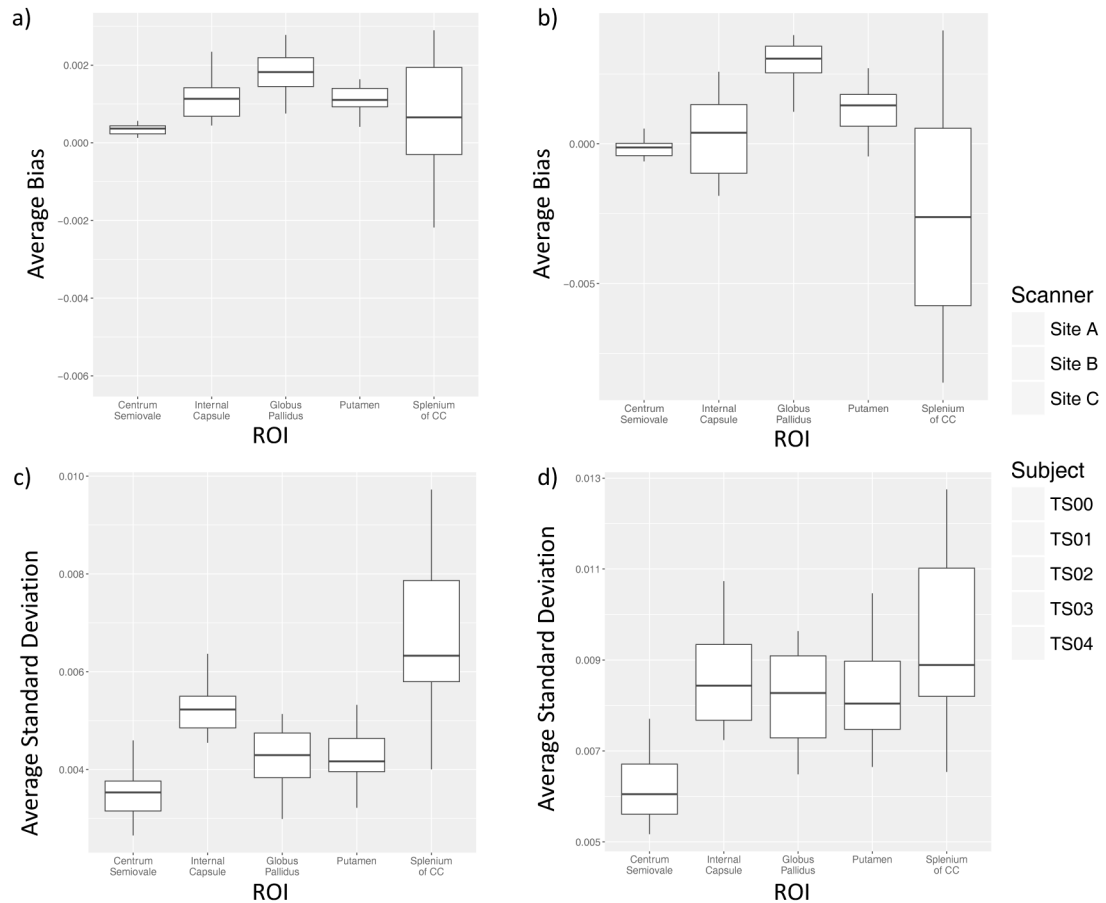
Figure 6.
Additional exploratory box plots for (a) average bias for b=1000 (b) average bias for b=2500 (c) average standard deviation for b=1000 (d) average standard deviation for b=2500. Each point represents the value averaged across all voxels within the ROI. Scan sites are identified by the shape of the data point and each subject is identified by a different color. The whiskers of the boxplot represent 1.5 times the inter-quartile range (IQR). Any points falling outside this range are considered outliers and should be investigated further. We find that subject 00 had a larger than expected negative bias value in the internal capsule for b=1000 which was the result of a segmentation error. In addition, several scans fell outside the IQR for the internal capsule on the b=2500 scans. Each scan identified as an outlier should be examined for artifacts or other errors that could impact the analysis and corrected before continuing.