Published in final edited form as:

Brain Imaging Behav. 2018 April ; 12(2): 437–448. doi:10.1007/s11682-017-9702-2.

An Evaluation of Z-Transform Algorithms for Identifying Subject-Specific Abnormalities in Neuroimaging Data

Andrew R. Mayer^{1,2,3,*}, Andrew B. Dodd¹, Josef M. Ling¹, Christopher J. Wertz¹, Nicholas A. Shaff¹, Edward J. Bedrick⁴, and Carlo Viamonte⁵

¹The Mind Research Network/Lovelace Biomedical and Environmental Research Institute, Albuquerque, NM 87106

²Neurology and Psychiatry Departments, University of New Mexico School of Medicine, Albuquerque, NM 87131

³Department of Psychology, University of New Mexico, Albuquerque, NM 87131

⁴Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ 85724

⁵Radiology Department, University of New Mexico School of Medicine, Albuquerque, NM 87131

Abstract

The need for algorithms that capture subject-specific abnormalities (SSA) in neuroimaging data is increasingly recognized across many neuropsychiatric disorders. However, the effects of initial distributional properties (e.g., normal versus non-normally distributed data), sample size, and typical preprocessing steps (spatial normalization, blurring kernel and minimal cluster requirements) on SSA remain poorly understood. The current study evaluated the performance of several commonly used z-transform algorithms [leave-one-out (LOO); independent sample (IDS); Enhanced Z-score Microstructural Assessment of Pathology (EZ-MAP); distribution-corrected z-scores (DisCo-Z); and robust z-scores (ROB-Z)] for identifying SSA using simulated and diffusion tensor imaging data from healthy controls (N=50). Results indicated that all methods (LOO, IDS, EZ-MAP and DisCo-Z) with the exception of the ROB-Z eliminated spurious differences that are present across artificially created groups following a standard z-transform. However, LOO and IDS consistently overestimated the true number of extrema (i.e., SSA) across all sample sizes and distributions. The EZ-MAP and DisCo-Z algorithms more accurately estimated extrema across most distributions and sample sizes, with the exception of skewed distributions. DTI results indicated that registration algorithm (linear versus non-linear) and blurring kernel size differentially affected the number of extrema in positive versus negative tails. Increasing the blurring kernel size increased the number of extrema, although this effect was much more prominent when a minimum cluster volume was applied to the data. In summary, current results

*Corresponding author: Andrew Mayer, Ph.D., The Mind Research Network, Pete & Nancy Domenici Hall, 1101 Yale Blvd. NE, Albuquerque, NM 87106; Tel: 505-272-0769; Fax: 505-272-8002; amayer@mrn.org.

Compliance with Ethical Standards

The authors declare that there are no conflicts of interest. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

highlight the need to statistically compare the frequency of SSA in control samples or to develop appropriate confidence intervals for patient data.

Keywords

simulations; single-subject; diffusion tensor imaging; neuroimaging; variability

Introduction

It is increasingly recognized that the number, nature and location of lesions changes during the typical course of most neurological diseases, including multiple sclerosis (Ge et al. 2005), vascular dementia (Rosenberg 2012) and external head trauma (Bigler et al. 2013). Subject-specific abnormalities (SSA) in neuroimaging data are also likely to be present in psychiatric disease states such as schizophrenia and substance abuse disorders (White et al. 2009). However, most neuroimaging studies focus on differences between *groups* of subjects rather than at the single-subject level, assuming that abnormalities consistently occur in identical locations (i.e., during voxel-wise analyses) or within the same region (i.e., during region of interest analyses). Unlike frank lesions that are visible on T₁, T₂, T₂^{*}, susceptibility-weighted imaging (SWI) and fluid-attenuated inversion recovery (FLAIR) sequences, diffusion (dMRI) and functional (fMRI) magnetic resonance imaging frequently depend on the utilization of statistical techniques for the detection of SSA (Mayer et al. 2014). However, the effects of initial distributional properties (e.g., normal versus skewed data), sample size and typical preprocessing steps (spatial normalization, registration, blurring, etc.) on the identification of SSA are poorly understood (Booth et al. 2016; Gebhard et al. 2015; Mayer et al. 2014).

The smoothing of neuroimaging data increases signal-to-noise ratio, improves between-subject comparisons, increases data normality and the Gaussian distribution of errors (Lindquist 2008; Parrish et al. 2000). However, visible lesions vary considerably in volume across disease states and individuals (Lucchinetti et al. 2000), and excessive smoothing can eliminate true SSA abnormalities (filter size > lesion; matched-filter theorem). Increasing the blurring kernel size also requires more stringent corrections to control for false positives (Lindquist 2008), which is usually accomplished by requiring a minimum threshold volume of contiguous voxels that is partially determined by data smoothness (Eklund et al. 2016; Friston et al. 1996; Hayasaka et al. 2004). To our knowledge, no studies have investigated how blurring kernel size affects the estimation of SSA. Recent SSA papers have highlighted the role of misregistration in the generation of false positives, the effects of subject-to-template or template-to-subject approaches, and how registration changing the initial shape of the distribution (Booth et al. 2016; Suri et al. 2015). Non-linear algorithms are also more sensitive than other commonly used techniques such as tract-based spatial statistics (TBSS) or linear registration for clinical comparisons using diffusion tensor imaging (DTI) data (Schwarz et al. 2014). However, the effect of registration algorithm on SSA results has also not been directly compared.

There is a family of SSA techniques that principally involve z-transforming data from both healthy controls (HC; hereafter referred to as reference group, RF) and patients (hereafter referred to as comparison group, CP) to determine the prevalence of SSA (Mayer et al. 2014). As a core principle, all methods z-transform RF and CP data using statistics derived from the reference group. SSA are then defined according to pre-determined statistical thresholds (e.g., $|z| > 2$) that approximate some percentile (e.g., cumulative 5th percentile across both tails) based on a *normal distribution*. Recent work and statistical theory have established that the variances of the reference (i.e., reduced) and comparison (i.e., increased) groups are different following a z-transformation (Bouix et al. 2013; Kim et al. 2013; Mayer et al. 2014; Watts et al. 2014), which can lead to a bias (i.e., more extrema in the comparison relative to reference group) if the two groups are directly compared. However, the distributional properties of neuroimaging data vary across modalities (i.e., fMRI versus DTI data), as well as within the same modality on a voxel-wise basis (Landman et al. 2012). Distributional properties are also affected by data preprocessing methods such as spatial blurring and registration (Booth et al. 2016; Gebhard et al. 2015; Mayer et al. 2014). However, the performance of different z-transform methods in non-normal distributions has not been examined.

The current study therefore extends our previous work (Mayer et al. 2014) in several important ways. First, we compared the performance of the four most popular z-transformation techniques (leave-one-out [LOO], independent sample [IDS], Enhanced Z-score Microstructural Assessment of Pathology [EZ-MAP] and distribution-corrected z-scores [DisCo-Z]) for identifying SSA in imaging studies on simulated data with several distributional properties (i.e., differing amounts of skew and kurtosis) rather than just on a normal distribution. These simulations are more likely to capture the range of initial distributions present in most types of neuroimaging data. Second, in contrast to previous work, we also evaluated whether the Robust Z-score transform (ROB-Z), which has been used in non-imaging fields to decrease the effects of skew (Birmingham et al. 2009), would both reduce group-wise bias (Bouix et al. 2013; Kim et al. 2013; Mayer et al. 2014; Watts et al. 2014) and provide a superior method for identifying SSA in skewed data. Third, the current study also investigated the effects of various preprocessing steps (registration algorithms, blur size and cluster requirements) on bias and the ability to identify SSA in DTI data derived from a sample of healthy participants.

2. Methods and Results

2.1 Algorithm Comparison on Simulated Data

Monte Carlo simulations evaluated how four commonly applied z-transformation methodologies (IDS, LOO, EZ-MAP and DisCo-Z) performed with known initial distributions and as a function of sample size. Synthetic datasets containing 147,244 points (corresponding to the number of 1 mm voxels in the John Hopkins University [JHU] white matter atlas) were randomly generated for a standard normal distribution ($\mu = 0$), a distribution with high kurtosis (6 degrees of freedom [df] *t* distribution), a distribution with high skew (6 df chi-square distribution) and a distribution with low skew (12 df chi-square). All of these distributions were also scaled to have a variance of 1. Monte Carlo simulations

were limited to chi-square distributions for the ROB-Z analyses. A bimodal distribution was also created to mimic regions that were more susceptible to registration errors (see Supplemental Materials) using a mixture of two normal distributions ($\mu_1 = 0.50$, variance = 0.0025, 70% of distribution; $\mu_2 = 0.30$, variance = 0.0025, 30% of distribution; value range = 0–1.0). For this specific simulation, values were purposefully selected to more closely mirror diffusion data (FA). For all distributions, each individual dataset was further constrained to have an average interclass-correlation of 0.10 through the addition of correlated noise, which roughly corresponded to the correlation measured from 500 randomly selected voxels in human DTI data (Mayer et al. 2014). Please see Supplemental Figure 1 for a graphical representation of each of the distributions.

Data were first z-transformed using the mean and standard deviation from the reference group for the IDS, LOO, EZ-MAP and DisCo-Z methods. For IDS, simulated data were generated for three different samples (a reference group [RF], a simulated HC study group and a simulated patient group [CP]). For the LOO method, each dataset in the RF group was individually z-transformed based on the mean and standard deviation from the remainder of the datasets in the sample, with the CP group z-transformed based on the entire RF sample (Bouix et al. 2013; Pasternak et al. 2014). A combination of methodologies was used for z-transforming data in the EZ-MAP method. First, similar to IDS, data from the HC study and patient groups were both z-transformed using a third reference sample. In addition, bootstrapping (2000 iterations) in combination with the LOO method was used to empirically determine the variance of z-scores in the reference sample (Kim et al. 2013). The calculated z-scores for the study RF (HC) and CP (patients) groups were then divided by the bootstrapped variance estimate (Kim et al. 2013). Finally, for all comparisons involving ROB-Z, data were transformed by the median and median absolute deviation (MAD) rather than the mean and standard deviation. Extrema were identified based on a set threshold (i.e., $|z| > 2$) for IDS, LOO, EZ-MAP and ROB-Z or based on adjusted z-scores that accounted for the distributional properties (i.e., DisCo-Z) of the transformed RF and CP samples (Mayer et al. 2014).

For all simulations, RF and CP samples were randomly selected with replacement from each distribution in samples sizes that ranged from 10 to 50 per group (CP, RF and third reference sample where applicable) in intervals of 10. A total of 400 samples were generated for each method and sample size. Our two principal measures of bias were 1) the total number of extrema in positive and negative tails (i.e., estimation bias) and 2) the number of samples where extrema statistically differed between the study RF and CP groups (i.e., group-wise bias). To empirically determine the total number of extrema in positive and negative tails, 4.0×10^7 data points were randomly sampled from each of the four distributions (Supplemental Figure 1) and z-transformed using the mean and standard deviation. A frequency count was used to empirically determine the expected ratio of extrema in each tail at $|z| > 2$. The large sample sizes possible in our simulation data eliminate the bias present in standard z-transformations at smaller sample sizes (Mayer et al. 2014). If group-wise bias is absent following data transformations, the proportion of samples where extrema statistically differ between the RF and CP sample should be equivalent to the predetermined alpha level. Average computational times for simulations involving the normal distribution are presented for LOO, IDS, EZ-MAP and DisCo-Z methods in Table 1.

Results—Simulations with a normal distribution indicated that all four methods negated group-wise bias (Figure 1a–d; second row) between the RF and CP groups, with the possible exception of the LOO method at the smallest sample size (RF > CP; N = 10 per group). However, both the IDS and LOO methods (Figure 1a and 1b) resulted in over-estimation of the number of extrema based on a standard normal distribution and the selected threshold of $|z| > 2.0$ (3353 extrema based on simulation). The over-estimation of extrema was dependent on sample size (i.e., greater at lower N) but was still evident at N = 50 per group for IDS and LOO methods. In contrast, the number of estimated extrema was more accurate and consistent as a function of N in both the EZ-MAP (Figure 1c) and DisCo-Z (Figure 1d) methods. Table 1 indicates computational times for various algorithms relative to DisCo-Z (i.e., the fastest method) at different sample sizes, with EZ-MAP requiring far greater computational resources relative to other methods.

Simulation results from the t distribution were similar both in terms of the elimination of group-wise bias (Figure 1e–h; second row) between the RF and CP samples for all methods and the over-estimation of extrema for IDS (Figure 1e) and LOO (Figure 1f) across all sample sizes. However, both the EZ-MAP (Figure 1g) and DisCo-Z (Figure 1h) methods also resulted in an over-estimation of extrema at lower N, which approached the expected value (~3673 based on simulations) as a function of increasing N.

All 4 (LOO, IDS, EZ-MAP and DisCo-Z) z-transformation methods also eliminated group-wise bias in chi-square distributions with either high (Figure 2a–d; second row) or low (Figure 2e–h; second row) positive skew. Over-estimation of extrema was present for IDS and LOO at both positive and negative tails of the distributions across all sample sizes. In contrast, EZ-MAP and DisCo-Z both approximated the predicted number of extrema in the positive (i.e., skewed) tail (high skew ~6505; low skew ~5869). However, both EZ-MAP and DisCo-Z also over-estimated extrema (expected levels: high skew ~0; low skew ~144) in the negative tail (see Supplemental Figure 2). To further investigate this finding, we also examined the accuracy of extrema estimation in larger sample sizes (N from 100 to 250 per group in intervals of 50) for the t and chi-square distributions using the DisCo-Z method. Results indicated that over-estimation of negative extrema was still present even at very large sample sizes for the chi-square distributions (Supplemental Figure 3). Similarly, none of the z-transform methods accurately identified the true number of extrema within either the positive or negative tail in the bimodal simulation (expected levels: positive ~149; negative ~3730; Supplemental Figure 4), with worse performance in the tail with few extrema.

We next utilized the ROB-Z transform to determine if performance would improve by using an algorithm specifically designed for robust outlier detection in skewed distributions. Unlike previous results with LOO, IDS, EZ-MAP and DisCo-Z, ROB-Z resulted in significant group-wise bias (CP > RF) across all sample sizes and in both high (Figure 3a) and low (Figure 3b) skew chi-square distributions, characterized by a high number of false positives when data were directly compared. Similar to previous results that simply used a traditional z-score transformation (reported in Mayer et al. 2014), the degree of bias was dependent on both sample size (bias greater at smaller N) and distributional characteristics (e.g., direction of skew). Interestingly, combining the DisCo-Z and ROB-Z resulted in over-

correction such that there was a higher false positive rate for the RF relative to CP group that was roughly equal across all sample sizes (Supplemental Figure 5).

In summary, simulation results indicated that all four z-transformation methods generally eliminated group-wise bias between two randomly selected samples regardless of initial distributional properties. Both IDS and LOO over-estimated the number of extrema in both positive and negative tails across all distributions, with the magnitude of over-estimation increasing as the sample size decreases. This results from the fact that normally distributed data assumes the shape of a t -distribution following a z-transformation, which has heavier tails at lower sample sizes (Mayer et al. 2014). Current results suggest that a similar bias exists across other distributions following LOO and IDS transformations, such that estimating the “expected” number of extrema is particularly challenging with these two algorithms.

ROB-Z, either alone or in concert with DisCo-Z, resulted in both estimation bias and group-wise bias. Thus, the ROB-Z does not appear to be suitable as a method for determining extreme values in neuroimaging data with skewed distributions. In contrast, the degree of overestimation was much smaller for both EZ-MAP and DisCo-Z for kurtotic distributions, although both methods over-estimated extrema in tails with few extrema (skewed or bimodal). Finally, results indicated that computational time was orders of magnitude shorter for DisCo-Z relative to EZ-MAP. As a result of these simulations, analyses examining the effect of various preprocessing steps on SSA were limited to the DisCo-Z algorithm.

2.2 Effects of Preprocessing Steps on Bias Metrics

Imaging data from 50 adult healthy controls (25 males; 27.42 ± 8.96 years old; 13.90 ± 2.09 years of education) were used to examine how various preprocessing steps impacted our two metrics of bias (i.e., group-wise and estimation) in real dMRI data. The data used for these analyses were identical to our previous publication (Mayer et al. 2014) and details are therefore only briefly presented here. All images were collected on a 3 Tesla Siemens Tim Trio scanner and included a high-resolution 5-echo T_1 MPRAGE sequence [TE = 1.64, 3.5, 5.36, 7.22, 9.08 ms, TR = 2.53 s, TI = 1.2 s, 7° flip angle, NEX = 1, slice thickness = 1 mm, FOV = 256×256 mm, voxel resolution = $1 \times 1 \times 1$ mm³]. Diffusion data were acquired using a twice-refocused spin echo sequence with 30 diffusion gradients ($b = 800$ s/mm²) and the $b = 0$ experiment repeated five times [72 interleaved slices; TE = 84 ms; TR = 9 s; 90° flip angle; NEX = 1; FOV = 256×256 mm; matrix size = 128×128 ; voxel resolution = $2 \times 2 \times 2$ mm]. The diffusion scan was repeated twice with identical parameters for a total of 60 diffusion weighted images.

Tissue segmentation of gray matter (GM), white matter (WM) and cerebral spinal fluid (CSF) was performed on each participant's T_1 data using SPM12. Raw DTI data and gradient tables were concatenated across the two runs and corrected for eddy currents and head motion using a 12 df affine registration. The vector corresponding to the rotation component was extracted from the resultant transformation matrix and applied to the gradient table. Prior to calculating diffusion tensors and scalar measures (fractional anisotropy; FA), images were smoothed anisotropically to improve signal-to-noise characteristics (Ding et al. 2005). A nonlinear method was adopted for tensor calculations to

decrease tensor estimate errors caused by noise, especially in regions of high anisotropy (Cox and Glen 2006). The first b-value = 0 image was registered to each participant's T₁ image using a 12 df affine transformation with Local Pearson Correlation as the cost function (Saad et al. 2009).

T₁ images were subsequently normalized to Talairach space (TT_N27 template from AFNI) using either a 12 df linear or a non-linear (3dQWARP from AFNI) transformation. The normalization matrix was subsequently concatenated with the diffusion to T₁ matrix, with the product of the two matrices used to normalize FA data into template space. A second factor examined the effect of blurring kernel size on SSA across the different registration pipelines. Specifically, FA data were blurred with a 3, 6 or 9 mm full-width at half maximum Gaussian kernel (GB) following both the linear and non-linear registration pipelines. The spatially normalized whole-brain FA maps from 50 HC were randomly sampled with replacement into either a RF or CP sample over 500 different iterations at N = 30 per group. All data were masked by the JHU white matter atlas from FSL (Mori and van Zijl 2007) to reduce the number of comparisons and restrict the analysis to WM. Data were analyzed both with and without a minimum cluster size threshold (128 μ l; 16 native voxels) to provide continuity with simulations and previous publications (Mayer et al. 2014). Measures of bias were identical to simulation data and included 1) the total number of extrema in positive and negative tails (Figure 4b & d) and 2) the number of iterations where extrema statistically differed between the reference and comparison groups (Figure 4a & c). See Supplemental Figure 6 for an example of a typical processing stream for imaging data involving DisCo-Z transformations.

Results—The number of significant differences between the RF and CP samples remained at chance levels across both linear and non-linear registration pipelines, blurring kernel sizes and thresholding strategies (Figure 4a & c), suggesting that none of the preprocessing steps increased the incidence of group-wise bias between the RF and CP samples. Two $2 \times 4 \times 2$ [Normalization (Linear vs. Non-linear) $\times$ Blur (0, 3, 6 and 9 mm) $\times$ Tail (Positive vs. Negative)] ANOVAs were performed to evaluate the effect of preprocessing steps on the number of extrema in the reference sample for both unthresholded and thresholded (minimum cluster size = 128 μ l) data. In these analyses, the mean number of extrema across the 30 randomly selected subjects on each of the 500 iterations served as the dependent variable.

For unthresholded data (Figure 4b), results from the ANOVA indicated significant main effects of registration ($F_{1,499} = 25.25$; $p < 0.001$; Non-linear > Linear), tail ($F_{1,499} = 16.04$; $p < 0.001$; Positive > Negative) and blur ($F_{3,1497} = 159.68$; $p < 0.001$). There was a small increase in the number of extrema as a function of blur kernel size (GB 9 [3312] > GB 6 [3274] > GB 3 [3219] > GB 0 [3191]), with the number of extrema approaching expectation (see dotted line in 4b) across all kernel sizes. There were also significant 2- and 3-way interactions across all combination of variables (all p 's < 0.001). As a result, only the Normalization $\times$ Blur $\times$ Tail interaction was analyzed further. Simple effects testing indicated that the difference in the number of extrema across the tails was slightly inverted for the different registration algorithms (linear = negative > positive extrema; non-linear = positive > negative extrema) for GB 0, 3 and 6 (all p 's < 0.001). At GB9, in contrast, the

direction of the tails was similar (positive > negative extrema) but of a different magnitude between registration algorithms ($p < 0.001$).

Results from the thresholded data (128 μ l; Figure 4d) again indicated a significant main effect for blur ($F_{3,1497} = 23548.86$; $p < 0.001$), with a more marked increase in the number of extrema as a function of blur kernel size relative to unthresholded data (GB 9 [3174] > GB 6 [2883] > GB 3 [2235] > GB 0 [1631]). Thus, the number of extrema were below expectation (see dotted line in 4d) for all kernel sizes below 9 mm. Main effects of registration ($F_{1,499} = 317.39$; $p < 0.001$; Linear > Non-linear) and tail ($F_{1,499} = 4.41$; $p = 0.04$; Negative > Positive) were also observed, along with significant 2- and 3-way interactions between variables (all p 's < 0.001). Simple effects testing indicated that the number of extrema was slightly inverted in the positive versus negative tail for the different registration algorithms (linear = negative > positive extrema; non-linear = positive > negative extrema) for GB 3 and 6 (all p 's < 0.001), with a similar direction (positive > negative extrema) but differing magnitude again observed at GB 9 ($p < 0.001$). The thresholded GB 0 data were characterized by a difference in tails for linear registration (negative > positive extrema) with a similar number of extrema for tails in the non-linear algorithm ($p < 0.001$).

Supplemental analyses (see Supplemental Figure 7) confirmed that a similar pattern of findings (i.e., directions and magnitudes of tail differences between algorithms) existed across both smaller ($N = 10$) and larger ($N = 50$) sample sizes with the exception of the main effect of tail (p 's > 0.10). Additional supplemental analyses (see Supplemental Materials) confirmed the superiority of non-linear relative to linear registration algorithms for dMRI data (Klein et al. 2009). Specifically, both decreased native smoothness and voxel-wise variance of FA (Supplementary Figure 8c) were observed for the non-linear relative to linear algorithm. FA values were also generally increased in white matter for the non-linear registration (Supplementary Figure 8b), which explains the increased positive skew of FA data observed in Figure 4b and 4d.

In summary, current analyses confirm that neither linear nor non-linear registration pipelines were associated with group-wise bias or over-estimation of extrema. The number of extrema increased with an increase in blur kernel size, although this effect was more prominent when a minimum cluster size was used to identify SSA. As expected, supplemental analyses confirmed that the non-linear algorithm was generally more accurate than the linear algorithm for the registration of dMRI data. Unexpectedly, the different registration algorithms affected the shape of FA distributions in WM (slight negative skew for linear registration versus slight positive skew for non-linear registration), with the degree of skew also changing as a function of blur kernel size.

Discussion

The first goal of the study was to investigate how different z-transformation methods (LOO, IDS, EZ-MAP, DisCo-Z, ROB-Z) performed at identifying SSA under different initial distributions (normal, kurtotic, low positive skew and high positive skew) and sample sizes (N ranging from 10–50 per group) common across neuroimaging studies. Importantly, relative to a standard z-transformation, LOO, IDS, EZ-MAP and DisCo-Z eliminated the

group-wise bias that occurs between reference and comparison groups at sample sizes greater than 20 (Mayer et al. 2014; Watts et al. 2014). In contrast, this bias was still present following the ROB-Z transformation, suggesting that the selection of different statistical metrics alone (median versus mean; MAD versus standard deviation) represents an insufficient correction.

Specifically, both the standard z transformation (Mayer et al. 2014) and ROB-Z transformation result in reduced variance in the reference group and increased variance in the comparison group, leading to a high number of false positives ($CP > RF$) when the two groups are directly compared (Mayer et al. 2014). Importantly, these biases are present regardless of whether data are normally or non-normally distributed, although distributional characteristics such as direction of skew are related to the magnitude of false positives (Mayer et al. 2014). Based on our implementation, simulations also clearly indicate reduced computational time for the DisCo-Z relative to other methods (i.e., $DisCo-Z < IDS < LOO < EZ-MAP$) due to a relatively efficient algorithm that simply requires a change in threshold values rather than multiple resampling steps (LOO and EZ-MAP). Although IDS was computationally much more efficient than LOO and returned similar results (i.e., corrected group-wise bias), the costly nature of MR data acquisition and the increased potential for the use of convenience samples renders this algorithm impractical from a feasibility perspective.

As expected, the number of extrema that were detected by the algorithms varied as a function of both initial distributional properties as well as sample size (Gebhard et al. 2015; Mayer et al. 2014). In general, the LOO and IDS algorithms performed similarly at sample sizes greater than $N = 20$, with both algorithms greatly overestimating the number of extrema across all distributions. Importantly, over-estimation of extrema was still present with IDS and LOO algorithms at $N = 50$ per group, a sample size larger than that typically employed in most neuroimaging studies and below the size needed to robustly detect SSA (Gebhard et al. 2015). When data are normally distributed initially, the z-transformed distribution of both the reference and comparison groups more closely resembles a t -distribution rather than a z-distribution with either the LOO or the IDS method. The t -distribution becomes more kurtotic (i.e., heavier tails) as a function of decreasing sample size, which artificially increases the number of observed extrema when a fixed threshold is employed based on *assumed properties* from a normal distribution (i.e., $|z|$ of 2 corresponds to 0.0228 of sample).

The EZ-MAP and DisCo-Z algorithms more closely approximated the true number of extrema in the normal distribution, for kurtotic distributions at higher sample sizes and for the positive tail of positively skewed distributions. Unlike LOO and IDS algorithms, both EZ-MAP and DisCo-Z correct for the changes in distributional variance that occur following a standard z-transformation, albeit through different methods (Kim et al. 2013; Mayer et al. 2014). Conversely, both EZ-MAP and DisCo-Z over-estimated the number of extrema for the negative tails of positively skewed distributions and for kurtotic distributions at small sample sizes ($N = 20$ per group) that are typical in most neuroimaging studies. However, the over-estimation of extrema was much smaller for EZ-MAP/DisCo-Z relative to IDS/LOO across all non-normal distributions, suggesting that the correction applied by both algorithms is more effective than not applying any correction at all. Recently, Gaussian error

propagation (Gebhard et al. 2015; Shaker et al. 2017) and a non-local means framework (Commowick and Stamm 2012) have been suggested as other corrective method for approximating SSA across multiple sample sizes and initial distributions.

We also investigated the effect of various preprocessing steps (registration algorithms, blurring kernel size and requirement of a minimal cluster size) on our two metrics of bias using FA data from a large sample of healthy controls. Monte Carlo simulations confirmed that none of the preprocessing steps critically affected the degree of group-wise bias following DisCo-Z correction. However, blurring kernel size and the interaction between kernel size and a minimum cluster requirement both affected the number of extrema observed across various preprocessing steps. Specifically, the number of extrema in healthy control DTI data more closely approximated expected levels when a minimum cluster threshold was not utilized (Figure 4b). The level of extrema remained relatively constant with increasing kernel sizes in spite of the increased covariance that occurs following the application of spatial blur. In contrast, requiring a minimal cluster size (128 μ l) decreased the number of observed extrema for all Gaussian kernels, with the largest effects present the smaller blur kernels (less than 6 mm). These results should readily generalize to other DTI metrics (e.g., mean/axial/radial diffusivity, mean kurtosis) given the similarity between their distributional properties and FA.

As previously discussed (Lindquist 2008; Parrish et al. 2000), spatial smoothing has several properties (i.e., increasing the signal-to-noise ratio and improving anatomical correspondence across subjects) that are important for SSA analyses due to the reliance on normative statistics. However, lesion volumes vary across different disease states (e.g., multiple sclerosis versus TBI), pathology types (e.g., a petechial hemorrhage versus a large contusion), and even within individuals (Lucchinetti et al. 2000). As such, it is difficult to determine an exact kernel size or minimum cluster requirement that is ideal for all types of SSA. Similarly, cluster-extent thresholding has been criticized due to its sensitivity to diffuse and potentially weak signals when using recommended defaults (Woo et al. 2014), with recent studies indicating inflated false positives for single-subject (Eklund et al. 2015) and group-wise (Eklund et al. 2016) fMRI results when using default clustering thresholds. Although false discovery rate (Booth et al. 2016), threshold-free cluster enhancement techniques (Jones and Cercignani 2010; Smith and Nichols 2009) and permutation testing (Nichols and Holmes 2002) have all been suggested for the correction of imaging data, to date no one technique has been fully verified as controlling for false positives both at the group and the single-subject level (Nichols and Hayasaka 2003). All of these factors highlight the need for rigorous statistical approaches for determining whether the detected number of SSA truly exceeds chance levels relative to a control sample, especially when sample sizes are modest (Booth et al. 2016; Gebhard et al. 2015; Mayer et al. 2014).

Not surprisingly, current results indicated that a non-linear versus linear registration algorithm resulted in 1) decreased smoothing of DTI data, 2) increased white matter FA and 3) decreased voxel-wise variance across all tissue types, all of which indicate the superiority of non-linear registration. Previous studies have also indicated the superiority of non-linear registration techniques when using either manual labeling as a gold standard or in evaluating the sensitivity/specificity of detecting clinical pathology (Ceritoglu et al. 2009; Klein et al.

2009; Schwarz et al. 2014). More germane to the current analyses, others groups have suggested that registering all other participants' data to a single participant's native space results in fewer SSA, which was interpreted to indicate that template based registrations produce false positives (Suri et al. 2015). Current results indicated an interaction between the type of registration algorithm employed and the observed number of extrema in positive (non-linear greater than linear) versus negative (linear greater than non-linear) tails at certain kernel sizes, suggesting that the choice of registration techniques alters the distribution (i.e., skew) of FA in white matter voxels. Thus, current and previous work highlights the important role that registration plays in any analyses focused on SSA, and strongly supports the utilization of nonlinear algorithms for template normalization.

Conclusions

In summary, it has been suggested that research-based imaging sequences (i.e., dMRI and fMRI) are more sensitive for detecting pathology in both animal and human studies across various neurological and psychiatric diseases (Commowick et al. 2008; Mac Donald et al. 2007). Robust statistical techniques are typically needed to quantify SSA in most neuropsychiatric populations due to the lack of a ground truth (i.e., a visible lesion). Current results suggest that both DisCo-Z and EZ-MAP methods provide a more accurate estimate of the true number of extrema relative to LOO and IDS. While it is unlikely that there will be a single method that is applicable to all of the different types distributions present in neuroimaging data, both the DisCo-Z and EZ-MAP methods performed well across all data simulations, suggesting that the correction is relatively robust. In contrast to the EZ-MAP, the DisCo-Z is also statistically straight-forward to implement and has relatively low computational costs based on our implementation of the algorithms.

Importantly, none of the methods accurately identified the true number of extrema for all distributions. Current results also suggest that the number of extrema was affected by several common preprocessing steps (registration algorithm, blurring kernel and cluster threshold) in FA data derived from healthy controls. Thus, there are many unforeseen consequences that occur during transformations and analyses of dMRI data (Booth et al. 2016; Jones and Cercignani 2010). Collectively, these results indicate that the sole reliance on a statistical threshold based on theoretical probabilities is not sufficient for SSA identification as the true underlying distribution is typically unknown in imaging data and also varies on a voxel-wise basis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health (grant numbers 1R01MH101512-01A1 and 1R01NS098494-01A1 to A.M.). The funding agencies had no involvement in the study design, data collection, analyses, writing of the manuscript, or decisions related to submission for publication. We would also like to thank Diana South and Catherine Smith for their assistance with data collection.

References

- Bigler ED, Abildskov TJ, Petrie J, Farrer TJ, Dennis M, Simic N, et al. Heterogeneity of brain lesions in pediatric traumatic brain injury. *Neuropsychology*. 2013; 27(4):438–451. [PubMed: 23876117]
- Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, et al. Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods*. 2009; 6(8):569–575. [PubMed: 19644458]
- Booth BG, Miller SP, Brown CJ, Poskitt KJ, Chau V, Grunau RE, et al. STEAM — Statistical Template Estimation for Abnormality Mapping: A personalized DTI analysis technique with applications to the screening of preterm infants. *Neuroimage*. 2016; 125:705–723. [PubMed: 26515903]
- Bouix S, Pasternak O, Rathi Y, Pelavin PE, Zafonte R, Shenton ME. Increased gray matter diffusion anisotropy in patients with persistent post-concussive symptoms following mild traumatic brain injury. *PLoS One*. 2013; 8(6):e66205. [PubMed: 23776631]
- Ceritoglu C, Oishi K, Li X, Chou MC, Younes L, Albert M, et al. Multi-contrast large deformation diffeomorphic metric mapping for diffusion tensor imaging. *Neuroimage*. 2009; 47(2):618–627. [PubMed: 19398016]
- Commowick O, Fillard P, Clatz O, Warfield SK. Detection of DTI white matter abnormalities in multiple sclerosis patients. *Med Image Comput Comput Assist Interv*. 2008; 11(Pt 1):975–982. [PubMed: 18979840]
- Commowick, O., Stamm, A. Non-local robust detection of DTI white matter differences with small databases. *Springer*; 2012. p. 476-484.
- Cox, R., Glen, D. Efficient, robust, nonlinear, and guaranteed positive definite diffusion tensor estimation. Seattle, WA. Proceedings of the International Society for Magnetic Resonance and Medicine, 14th Scientific Meeting; 2006.
- Ding Z, Gore JC, Anderson AW. Reduction of noise in diffusion tensor images using anisotropic smoothing. *Magnetic Resonance Medicine*. 2005; 53(2):485–490.
- Eklund A, Nichols TE, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A*. 2016; 113(28):7900–7905. [PubMed: 27357684]
- Eklund, A., Nichols, T., Andersson, M., Knutsson, H. Empirically investigating the statistical validity of SPM, FSL and AFNI for single subject fMRI analysis. *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*; 2015. p. 1376-1380.
- Friston KJ, Holmes A, Poline JB, Price CJ, Frith CD. Detecting activations in PET and fMRI: Levels of inference and power. *Neuroimage*. 1996; 4(3 Pt 1):223–235. [PubMed: 9345513]
- Ge Y, Law M, Grossman RI. Applications of diffusion tensor MR imaging in multiple sclerosis. *Ann N Y Acad Sci*. 2005; 1064:202–219. [PubMed: 16394158]
- Gebhard, T., Koerte, I., Bouix, S. Sample size estimation for outlier detection. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Switzerland: Springer International Publishing; 2015. p. 743-750.
- Hayasaka S, Phan KL, Liberzon I, Worsley KJ, Nichols TE. Nonstationary cluster-size inference with random field and permutation methods. *Neuroimage*. 2004; 22(2):676–687. [PubMed: 15193596]
- Jones DK, Cercignani M. Twenty five pitfalls in the analysis of diffusion MRI data. *NMR in Biomedicine*. 2010; 23(7):803–820. [PubMed: 20886566]
- Kim N, Branch CA, Kim M, Lipton ML. Whole brain approaches for identification of microstructural abnormalities in individual patients: Comparison of techniques applied to mild traumatic brain injury. *PLoS ONE*. 2013; 8(3):e59382. [PubMed: 23555665]
- Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*. 2009; 46(3):786–802. [PubMed: 19195496]
- Landman, B.A., Yang, X., Kang, H. Do we really need robust and alternative inference methods for brain MRI?. *Springer*; 2012. p. 77-93.
- Lindquist MA. The statistical analysis of fMRI data. *Statistical Science*. 2008; 23(4):439–464.

- Lucchinetti C, Bruck W, Parisi J, Scheithauer B, Rodriguez M, Lassmann H. Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. *Ann Neurol.* 2000; 47(6): 707–717. [PubMed: 10852536]
- Mac Donald CL, Dikranian K, Bayly P, Holtzman D, Brody D. Diffusion tensor imaging reliably detects experimental traumatic axonal injury and indicates approximate time of injury. *J Neurosci.* 2007; 27(44):11869–11876. [PubMed: 17978027]
- Mayer AR, Bedrick EJ, Ling JM, Toulouse T, Dodd A. Methods for identifying subject-specific abnormalities in neuroimaging data. *Hum Brain Mapp.* 2014; 35(11):5457–5470. [PubMed: 24931496]
- Mori S, van Zijl PC. Human white matter atlas. *Am J Psychiatry.* 2007; 164(7):1005. [PubMed: 17606649]
- Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res.* 2003; 12(5):419–446. [PubMed: 14599004]
- Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp.* 2002; 15(1):1–25. [PubMed: 11747097]
- Parrish TB, Gitelman DR, LaBar KS, Mesulam MM. Impact of signal-to-noise on functional MRI. *Magn Reson Med.* 2000; 44(6):925–932. [PubMed: 11108630]
- Pasternak O, Koerte IK, Bouix S, Fredman E, Sasaki T, Mayinger M, et al. Hockey Concussion Education Project, Part 2. Microstructural white matter alterations in acutely concussed ice hockey players: a longitudinal free-water MRI study. *J Neurosurg.* 2014; 120(4):873–881. [PubMed: 24490785]
- Rosenberg GA. Neurological diseases in relation to the blood-brain barrier. *J Cereb Blood Flow Metab.* 2012; 32(7):1139–1151. [PubMed: 22252235]
- Saad ZS, Glen DR, Chen G, Beauchamp MS, Desai R, Cox RW. A new method for improving functional-to-structural MRI alignment using local Pearson correlation. *Neuroimage.* 2009; 44(3): 839–848. [PubMed: 18976717]
- Schwarz CG, Reid RI, Gunter JL, Senjem ML, Przybelski SA, Zuk SM, et al. Improved DTI registration allows voxel-based analysis that outperforms tract-based spatial statistics. *Neuroimage.* 2014; 94:65–78. [PubMed: 24650605]
- Shaker M, Erdogmus D, Dy J, Bouix S. Subject-specific abnormal region detection in traumatic brain injury using sparse model selection on high dimensional diffusion data. *Medical Image Analysis.* 2017
- Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage.* 2009; 44(1):83–98. [PubMed: 18501637]
- Suri AK, Fleysher R, Lipton ML. Subject based registration for individualized analysis of diffusion tensor MRI. *PLoS ONE.* 2015; 10(11):e0142288. [PubMed: 26580077]
- Watts R, Thomas A, Filippi CG, Nickerson JP, Freeman K. Potholes and molehills: Bias in the diagnostic performance of diffusion-tensor imaging in concussion. *Radiology.* 2014; 272(1):217–223. [PubMed: 24635677]
- White T, Schmidt M, Karatekin C. White matter ‘potholes’ in early-onset schizophrenia: a new approach to evaluate white matter microstructure using diffusion tensor imaging. *Psychiatry Res.* 2009; 174(2):110–115. [PubMed: 19853414]
- Woo CW, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage.* 2014; 91:412–419. [PubMed: 24412399]

Fig. 1.

Simulations comparing biases in the independent sample (IDS; Columns a and e), leave-one-out (LOO; Columns b and f), EZ-MAP (Columns c and g) and the DisCo-Z (Columns d and h) algorithms. Data were initially derived from either a normal distribution or a t distribution with six degrees of freedom (df) for both reference (RF) and comparison (CP) groups.

Sample size (N) within each group is plotted on the x-axis. The first and third rows present the mean number of extreme voxels that were observed in each method. Error bars depict the average standard error of the mean across iterations, while the horizontal dotted line represents the empirically determined number of extrema for $z > |2|$. The second and fourth rows depict group-wise bias, measured by the percentage of t-tests that were statistically different ($p < 0.05$) between the RF and CP groups (LOO and DisCo-Z) or two CP (CP1 vs. CP2; IDS and EZ-MAP) groups. Data are color-coded to reflect the number of extrema (Rows 1 and 3) or t-test results (Rows 2 and 4) from the negative (Neg; black and light gray bars) and positive (Pos; dark gray and white bars) tails. All methods eliminated group-wise

bias associated with unequal variances following z-transformations, whereas accuracy for estimating the expected number of extrema was more variable

Chi-Square Distribution (df = 6)

Fig. 2.

Simulations comparing bias in the independent sample (IDS; Columns a and e), leave-one-out (LOO; Columns b and f), EZ-MAP (Columns c and g) and DisCo-Z (Columns d and h) algorithms. Data were derived from high (Chi-Square Distribution: 6 degrees of freedom [df]) and low (Chi-Square Distribution: 12 df) positively skewed distributions at various sample sizes (N) for both reference (RF) and comparison (CP) groups. As in Figure 1, the first and third rows present the number of extrema whereas the second and fourth rows depict group-wise bias. Error bars depict the average standard error of the mean across iterations, while the horizontal dotted line represents the empirically determined number of extrema for $z > 2$. The number of expected negative extrema ($z < -2$) was either zero (df = 6) or near zero (144; df = 12) for high and low (see Supplemental Figure 2) skew distributions. Data are color-coded to reflect the number of extrema (Rows 1 and 3) or t-test results (Rows 2 and 4) from the negative (Neg; black and light gray bars) and positive (Pos; dark gray and white bars) tails. All methods eliminated group-wise bias associated with

unequal variances following z-transformations, whereas accuracy for estimating the expected number of extrema was more variable

Robust Z Simulations

Fig. 3.

Simulations exploring biases in the Robust Z method. Data were initially derived from a Chi-Square distribution with either a high (6 degrees of freedom [df]; Panel a) or low (12 df; Panel b) positive skew for both reference (RF) and comparison (CP) groups. Sample size (N) per group is plotted on the x-axis. The first row presents the mean number of extreme voxels that were observed in each distribution. Error bars depict the average standard error of the mean across iterations, while the dotted line represents the empirically determined number of extrema for $z > 2$. The second row depicts group-wise bias, measured by the percentage of t-tests that were statistically different ($p < 0.05$) between the RF and CP groups. Data are color-coded to reflect the number of extrema (Panel a) or t-test results (Panel b) from the negative (Neg; black and light gray bars) and positive (Pos; dark gray and white bars) tails.

Use of the Robust Z method led to an overestimate of extrema for the CP group, leading to a large number of false positive t-tests, especially at low N

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 4.

Fractional anisotropy (FA) data from 50 healthy controls who were randomly sampled with replacement into reference (RF) and comparison (CP) groups across 500 iterations ($N = 30$ per group). FA data were normalized to template space using either a linear or non-linear algorithm and blurred using full-width half-maximum Gaussian (GB) kernels of either 3, 6 or 9 mm or not blurred (GB 0). Data are presented both without any minimum cluster size threshold (Panels a and b) and with a threshold of 128 microliters (μl ; Panels c and d).

Horizontal dotted line in Panels b and d represent the empirically determined number of extrema for $z > |2|$. There was no evidence of systematic differences (false positives; FP) between the RF and CP groups as a function of thresholding, registration or blur size (Panels a and c) in either the negative (Neg; black and light gray bars) or positive (Pos; dark gray and white bars) tails of the distribution. The number of extreme voxels ($z > |2|$) across positive and negative tails was affected by the two registration pipelines and the kernel size.

Table 1

Average computation time (CT) of z-transform methods using a normal distribution with simulated data.

Sample Size (N)	IDS	LOO	EZ-MAP	DisCo-Z
10	0.15(0.02) s	0.22(0.03) s	31.29(2.08) s	0.08(0.01) s
20	0.25(0.00) s	0.68(0.01) s	50.19(1.79) s	0.18(0.01) s
30	0.36(0.00) s	1.21(0.01) s	71.18(0.50) s	0.25(0.00) s
40	0.48(0.00) s	1.99(0.03) s	89.65(0.53) s	0.34(0.02) s
50	0.62(0.03) s	2.91(0.02) s	104.03(0.49) s	0.42(0.03) s
CT	1.47	6.88	246.25	1

Notes: Mean and standard deviations provided from 100 iterations. N equals sample sizes for both reference and comparison groups as appropriate.
CT = computation time factor for all methods relative to DisCO-Z analyses at N = 50.