Original Research

Evaluation of Performance Metrics for Bias Field Correction in MR Brain Images

Zin Yan Chua, BS,[1] Weili Zheng, PhD,[1] Michael W.L. Chee, MBBS,[2] and Vitali Zagorodnov, PhD[1]*

Purpose: To investigate inconsistencies between common performance measures for bias field correction reported in several recent studies and propose a solution.

Materials and Methods: A set of synthetic images of a normal brain from the Montréal Simulated Brain Database (SBD) was processed using two bias field correction algorithms. The parameters of these algorithms were varied and the resulting outputs were assessed using several performance measures. Validity was estimated using Spearman rank correlation coefficient between "indirect" performance measures and the L2 norm of the difference between true and estimated bias fields. The "indirect" performance measures tested were: coefficients of variation of white matter (WM) and gray matter (GM), coefficient of joint variation. These measures were tested on bias field-corrected images that were permuted in terms of quality of WM/GM segmentation as well as the presence or absence of light smoothing.

Results: Existing indirect performance measures yielded poor validity scores, explaining the inconsistencies reported in the literature. Image noise and inappropriate inclusion of partial volume voxels and neighboring tissues were found to be contributory. Combining conservative segmentation and smoothing significantly improved validity.

Conclusion: The use of indirect performance measures in the conventional manner to guide bias field correction is unreliable. Using these metrics on lightly smoothed images with conservatively segmented tissues proved more reliable for guiding the selecting of parameters for nonuniformity correction ultimately contributing to more accurate brain segmentation.

Key Words: intensity nonuniformity correction; performance metrics; coefficient of variation; coefficient of joint variation

J. Magn. Reson. Imaging 2009;29:1271–1279.

THE PRESENCE OF intensity nonuniformity in images obtained from high-field magnetic resonance (MR) systems may adversely affect qualitative and quantitative image analysis. Intensity nonuniformity is characterized by the occurrence of a smoothly varying and multiplicative intensity field, also referred to as a bias field. Poor radiofrequency (RF) coil design, gradient-eddy currents, local variations in flip angle, and subject–scanner interactions are contributory (1). While some of these factors can be dealt with in hardware, for example, "central brightening" can be reduced with a multichannel phased-array receiver coil (2), a degree of retrospective correction is often beneficial. There exist detailed reviews of recent retrospective nonuniformity correction algorithms (1,3–5).

As there are many performance validation measures, selecting the best correction algorithm can be difficult. Existing correction measures can be categorized into three groups (5):

1) *Measures comparing true and estimated bias fields.* These include correlation (6), root mean square error (RMS) (7–9), standard deviation error (STD) (10), mean square error (MSE), and mean square distance (MSD) (4,11). The application of these measures is usually limited to simulated images, as the true bias field contained in actual MR images is unknown.

2) *Measures based on intensity variability.* These rely on the fact that bias field increases intensity variation within each tissue and assume that there is no change in noise level or scaling in mean intensity across tissues. Popular methods in this category include coefficient of variation of white matter (CV_{WM}), coefficient of variation of gray matter (CV_{GM}) (3,10,12–18), and coefficient of joint variation (*CJV*) between WM and GM (3,10,14,15). Minimizing *CJV* results in minimizing intensity variability within each tissue, while maintaining a good separation between mean tissue intensities. This circumvents the problem of poor segmentation performance, consequent on increase in overlap between in-

[1]School of Computer Engineering, Nanyang Technological University, Singapore.

[2]Cognitive Neuroscience Laboratory, Duke-NUS Graduate Medical School, Singapore.

Contract grant sponsor: A*STAR, Singapore (Agency for Science and Technology and Research); Contract grant number: SBIC C-012/2006.

*Address reprint requests to: V.Z., School of Computer Engineering, Nanyang Technological University, 50 Nanyang Ave., Singapore, 639798. E-mail: zvitali@ntu.edu.sg

Received October 2, 2008; Accepted February 10, 2009.

DOI 10.1002/jmri.21768

Published online in Wiley InterScience (www.interscience.wiley.com).

Table 1
Excerpt of Performance Evaluation Results From Table II of Ref. 14 (1.5 T Scanner)

	CV_{WM}	CV_{GM}	CJV
SPM			
99-S	5.53	16.31	93.40
N3	5.88	16.47	83.95
NIC	5.84	15.70	80.77

tensity distributions. Both *CV* and *CJV* minimally require a coarse (but conservative) tissue labeling for WM and GM.

3) *Measures based on segmentation performance.* These use the quality of subsequent segmentation as a marker of the bias field correction performance, and include false-positive (FP) and false-negative (FN) rates (4), misclassification ratio (MCR) (19–21), Jaccard Similarity (JS) (18,22), and Dice coefficient (20,23,24).

The metrics belonging to the last two categories can be referred to as *indirect*, as the quality of the nonuniformity correction is derived indirectly through tissue intensity variability or segmentation performance. A survey of the recent literature revealed that such metrics often lead to conflicting suggestions regarding a best-performing method. For example, consider the data in Table 1, excerpted from a performance evaluation study (14), that compared three approaches (SPM99-S, N3, and NIC) using three indirect measures (CV_{WM}, CV_{GM}, and *CJV*). On the basis of CV_{WM}, SPM99-S (25) was the best-performing method. However, it was the worst performer if *CJV* was used. According to CV_{GM} and *CJV*, the best-performing method was NIC (14). The N3 correction algorithm (17) was the worst-performing according to CV_{WM} and CV_{GM} but second best according to *CJV*. Similar inconsistencies have been reported in other studies (3,10,15).

Given that a direct measurement of the bias field in experimentally collected human subject MR data is not feasible, it is important to determine the validity of the indirect metrics, ie, their ability to consistently reflect the quality of the true bias field correction, characterized by direct metrics. Here we established a link between direct and indirect measures using a simulated dataset in which the bias field is known a priori. Our experiments show that the existing measures may exhibit poor validity, particularly when the underlying WM/GM segmentation is subpar. Conservative labeling of WM/GM can increase validity and this can be incremented further by slight smoothing of the image data.

MATERIALS AND METHODS

Modeling Intensity Nonuniformity

The intensity nonuniformity is usually modeled using a smooth multiplicative field (1,5,13,16,19,21):

$$I(s) = B_{true}(s)I_0(s) + n(s) \quad [1]$$

Here s denotes a voxel, $I(s)$ and $I_0(s)$ are the intensities of corrupted and ideal (without noise and intensity nonuniformity) images, respectively. $B_{true}(s)$ denotes the bias field and $n(s)$ is the image noise. This model is consistent with RF field mapping theory that links voxel intensities with RF coil transmission and reception sensitivity.

The goal of bias field correction is to estimate the unknown $B_{true}(s)$ given $I(s)$, and to then use $B_{true}(s)$ to obtain an intensity corrected image. Without additional constraints this problem is inherently ill-posed, as there can be infinite combinations of $B_{true}(s)$ and $\frac{I_0(s)}{0}$, all giving rise to the same product, $B_{true}(s)\mathrm{I_o}$. To obtain a unique solution, most current methods regularize the problem by assuming that the bias field is spatially smooth and that tissue intensities fall into a finite set of discrete classes (GM, WM, and cerebrospinal fluid [CSF]), or that the peaks of intensity distribution are sharpest when nonuniformity has been corrected (9). These assumptions do not hold exactly for images obtained from human subjects, as a result of partial volume effects or biologically driven regional differences in image intensity (26), yet they underpin the majority of current methods.

Performance Measures Tested

For a direct measure, we used the normalized L_2-norm of the difference between the true $B_{true}(s)$ and the estimated $B_{est}(s)$ bias fields, defined as:

$$L_2 = \min_w \sqrt{\frac{\sum_s (wB_{est}(s) - B_{true}(s))^2}{\sum_s B_{true}^2(s)}} \quad [2]$$

where w is the normalization coefficient. Normalization is necessary because nonuniformity correction can result in an arbitrary scaling of the bias field. It is straightforward to show that the coefficient w must satisfy:

$$w = \frac{\sum_s B_{true}(s)B_{est}(s)}{\sum_s B_{est}{}^2(s)} \quad [3]$$

For the indirect measures, we used common *CV* and *CJV*, which are defined as:

$$CV_T = \frac{\sigma(T)}{\mu(T)}, \quad CJV = \frac{\sigma(WM) + \sigma(GM)}{|\mu(WM) - \mu(GM)|} \quad [4]$$

where T is a single tissue class (WM or GM) and $\sigma(T)$, $\mu(T)$ denote the standard deviation and the mean intensity within T, respectively. In this context, smaller *CV* and *CJV* correspond to smaller remaining bias field and hence better correction performance.

Measuring the Validity of Indirect Measures

Validity, in a statistical sense, refers to the consistency between a measurement and a criterion. In the context of nonuniformity correction, measurements are repre-

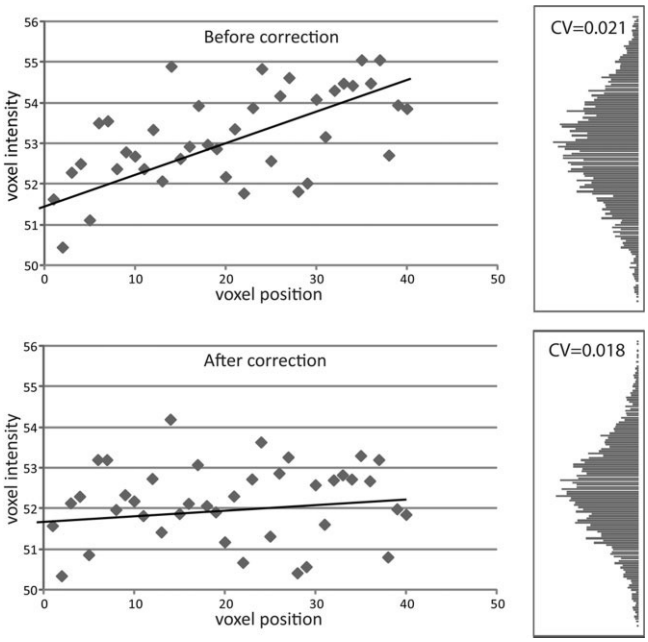

Figure 1. The link between tissue variability and intensity nonuniformity.

sented by indirect measures and the criterion by direct measures. Consistency is defined as a monotonic relationship between the measures. An indirect measure γ is perfectly consistent with the direct measure d, if for any two corrected results c_1 and c_2, $d(c_1) > d(c_2)$ implies $\gamma(c_1) > \gamma(c_2)$. None of the issues raised in the introduction would arise if indirect measures behaved in this manner.

We chose Spearman rank correlation coefficient ρ as a metric for validity. Spearman correlation between two sets of data X_i and Y_i is defined as the Pearson product-moment coefficient in which X_i and Y_i are converted to rankings x_i and y_i (27). Perfect consistency between indirect and direct measures implies a preservation of rank order, hence $\rho = 1$. Violations of consistency for at least some data points would lead to a mismatch between rankings and a reduction of ρ.

Factors Contributing to the Reduced Validity of Indirect Measures

To examine the factors that contribute to reduced consistency between tissue intensity variability and nonuniformity we used a simple simulation. The bias field was modeled using a linear function and image noise was generated using the MatLab (MathWorks, Natick, MA) function "randn" (Figs. 1, 2). Here the intensity variability can be visually gauged by the width of the histogram provided on the right side of the plots. As shown in the upper plot of Fig. 1, both image noise and bias field contribute to variability. The correction reduces or eliminates the bias field, resulting in reduced intensity variability, as shown in the lower plot (Fig. 1). However, when several data points from a different tissue were erroneously included in the reference tissue, the postcorrection variability actually increased by almost 10% (from 0.022 to 0.024; see Fig. 2). This was caused by nonuniformity-caused intensity overlap between the darker (smaller amplitude) points of the reference tissue and the wrongly included data points. The correction eliminated this overlap, producing a flatter left tail in the histogram (caused by wrongly included voxels) and a concomitant increase in variability.

These observations suggest two causes for the decrease in validity of indirect measures: nonideal tissue segmentation and image noise. To test the first part of this hypothesis we evaluated indirect measures under three types of underlying GM/WM segmentation:

1) *Ideal segmentation*. Perfect labeling of GM and WM tissues, taking into account partial volume voxels (3,4,10,15).

2) *Conservative segmentation*. Ideal segmentation followed by exclusion of partial volume voxels (9,28), implemented using morphological one-voxel deep erosion of each tissue class. Conservative segmentation can also be achieved manually (18).

3) *Corrupted segmentation*. Segmentation obtained by intentionally introducing misclassified voxels into ideal segmentation to model potential errors that may arise during expert-guided or automatic segmentation. The corruption was achieved by adding Gaussian noise to each binary class label with subsequent thresholding and removal of disconnected voxels.

To test the second part of the hypothesis, we included a novel set of indirect measures (referred to as modified *CV* and *CJV*), which were applied on slightly smoothed image data. The smoothing was achieved by replacing the intensity value of each voxel with the mean intensity of the $3 \times 3 \times 3$ voxel cube enclosing it. Smoothing was restricted to individual tissue classes to avoid averaging across tissue boundaries. We expected this small smoothing kernel to reduce variability due to image noise, without significantly influencing the shape of the bias field (assumed to be very smooth).

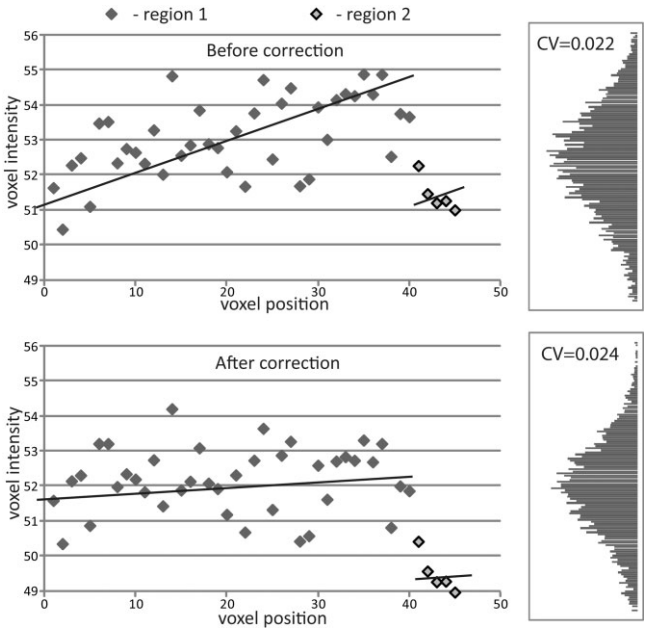

Figure 2. Failure of the coefficient of variation when a different tissue region is included.

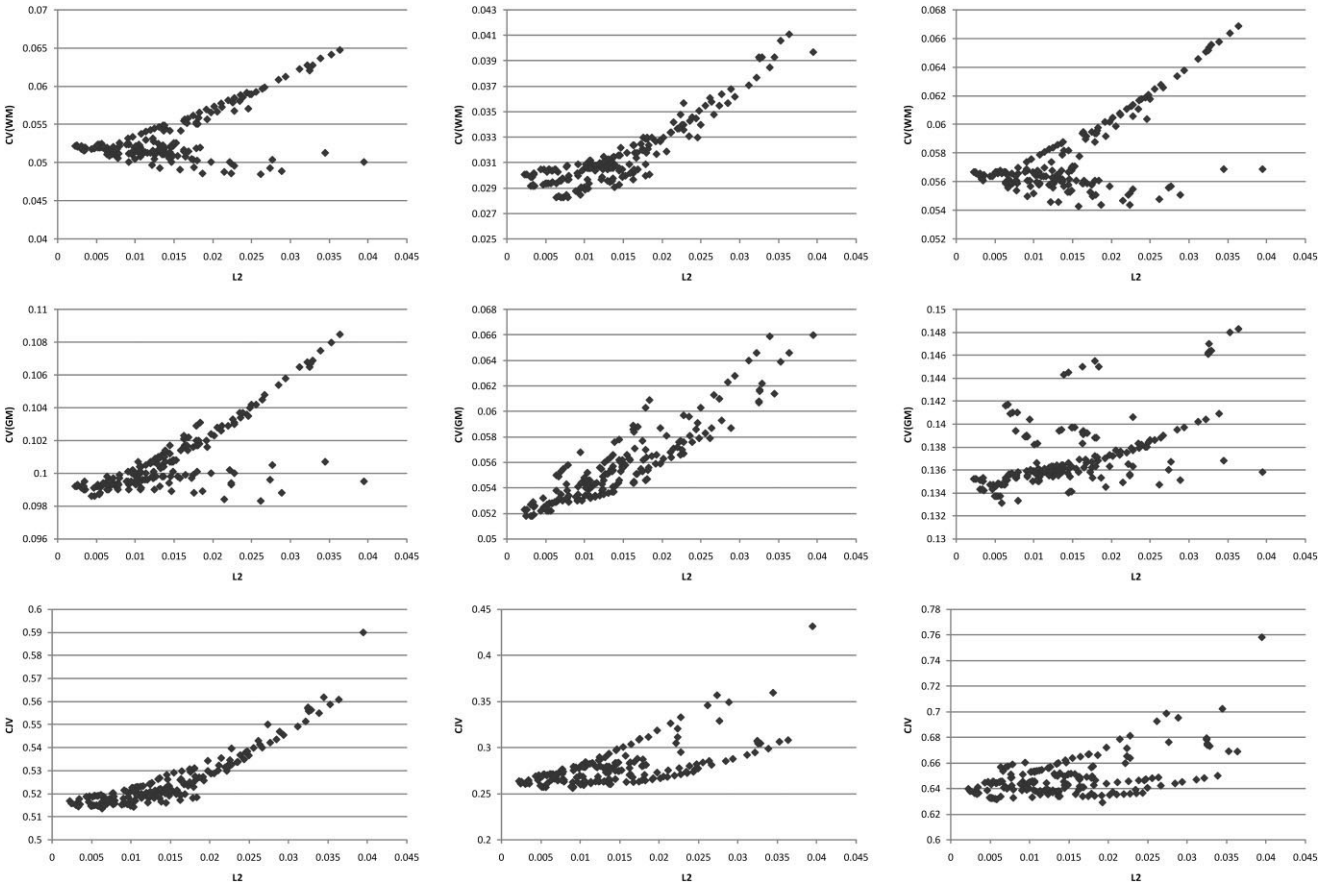

Figure 3. Scatterplots of CV_{WM}, CV_{GM}, and CJV versus L_2-norm using ideal (a), conservative (b), and corrupted (c) WM/GM segmentations.

Data and Experiment Setup

We used synthetic $1 \times 1 \times 1$ mm resolution T1-weighted MR data from the Montréal Simulated Brain Database (29–31) that was designed to simulate images of a normal brain, corrupted by various degrees of noise and intensity nonuniformity. In total, nine volumes were used, including all possible combinations of 0%, 20%, and 40% bias fields and 1%, 3%, and 5% noise. Exact tissue labeling (WM, GM, and CSF), provided by the dataset, was used for evaluation of indirect measures.

The experimental setup was as follows: each volume in the dataset was corrected for intensity nonuniformity by two algorithms, producing a set of corrected volumes together with their estimated bias fields. The direct measure [2] was then applied to the estimated bias fields, comparing them with the ground truth, while the three indirect metrics [4] were applied to the corresponding corrected volumes.

The two bias field correction algorithms correspond to two approaches: histogram sharpening (4,15,17,19,22) and surface fitting (3,13,16,32). The first algorithm was implemented on the basis of N3 (17), which iterates between sharpening of the image histogram (by deconvolving it with a Gaussian kernel), using the sharpened histogram to estimate the bias field at each voxel location, and imposing smoothness on the bias field. To obtain different correction samples, the smoothing kernel size was varied from 30 mm to 200 mm, in 10-mm increments. This was the key parameter pertaining to the amount of correction required (9). Smaller kernel sizes may result in a better fit to the true bias field, but can cause overfitting to noise and small image structures. Larger kernel sizes afford more conservative correction. The brain mask also exerts a substantial influence on the performance of this approach (12). Consequently, correction was applied to images with and without a brain mask, resulting in a total of $18 \times 2 \times 9 = 324$ different image sets.

The second algorithm (surface fitting) was implemented as follows. First, a coarse portion of the WM region was estimated using a standard region-growing procedure. Implementation began with a seed point, chosen automatically within the WM, and iteratively expanded into neighboring voxels based on intensity similarity. The parameters of the region-growing algorithm were chosen so that the resultant region conservatively estimated the WM. The intensities of the voxels within the estimated portion were then modeled using a linear combination of smooth-varying polynomials up to the third degree. The model coefficients were estimated using the method of least-squares. The performance of this approach is dependent on the highest polynomial degree used (higher degrees lead to sharper estimated bias fields and better fitting) and the size of the WM region (lager regions are less sensitive to noise but may overgrow into neighboring image structures).

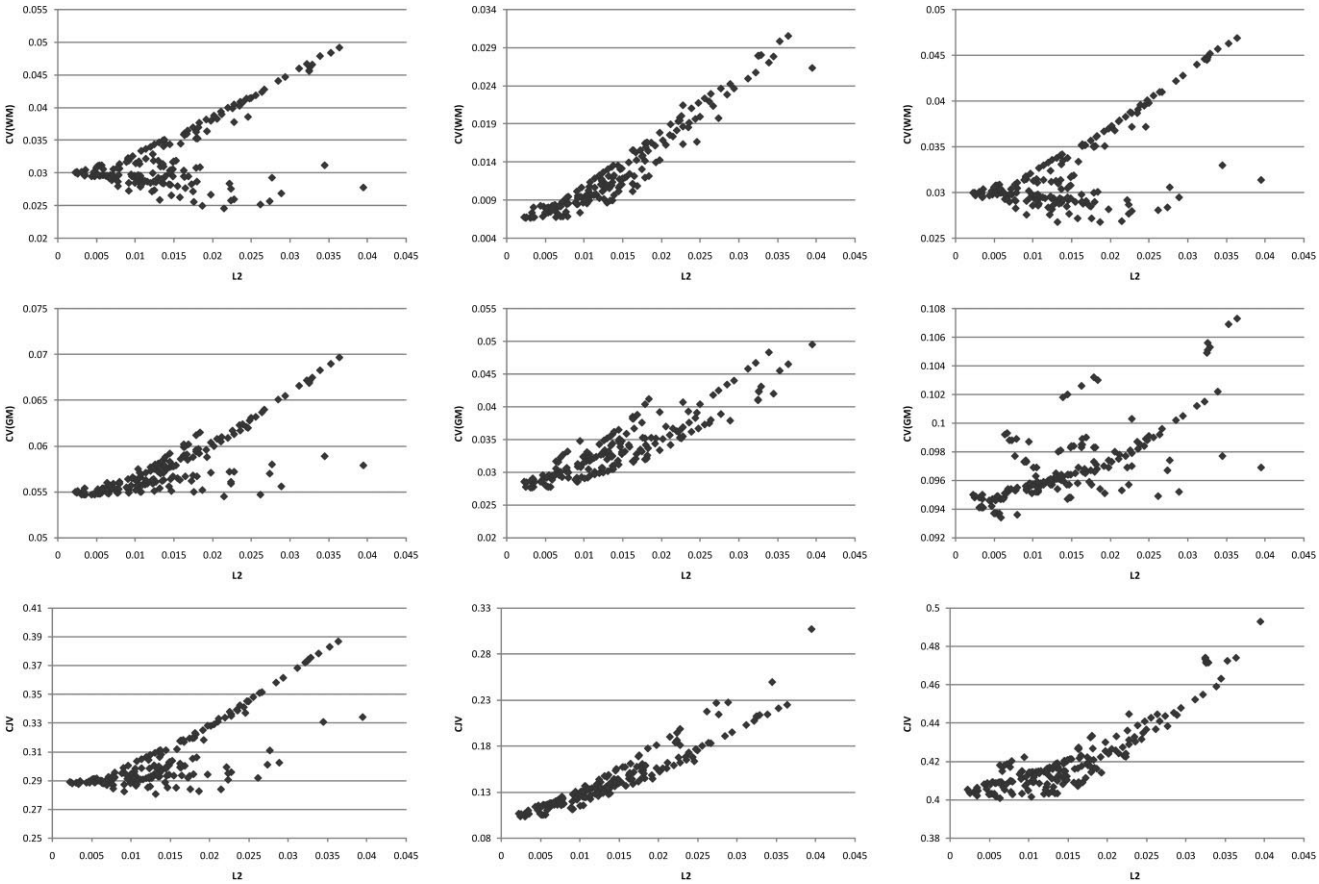

Figure 4. Scatterplots of modified CV_{WM}, CV_{GM}, and CJV versus L_2-norm using ideal (**a**), conservative (**b**), and corrupted (**c**) WM/GM segmentations.

In order to obtain different samples, region-growing parameters were manipulated. The generated regions from 5% to 60% of the WM were ramped up in 5% increments. The highest-order polynomial used varied, from 1 to 3. This process produced $12 \times 3 \times 9 = 324$ corrected images and corresponding estimated bias fields.

RESULTS

Qualitative Results

Figures 3 and 4 show the scatterplots between all indirect metrics and L_2-norm using ideal, conservative, and corrupted WM/GM segmentations. Here we combined results from the volumes with 3% noise and three bias field strengths (0%, 20%, and 40%).

The plots confirmed our expectation that segmentation quality can significantly affect the validity of indirect metrics. Overall, using conservative segmentation produced the most tightly clustered plots, the only exception being *CJV* evaluated using ideal segmentation. Preserving partial volume voxels (as in ideal segmentation) had a detrimental effect on CV_{WM} and CV_{GM}, increasing the scatter, with a further decrease in quality for all three metrics when the segmentation was corrupted. In particular, some portions of the scatterplots had negative slopes, reversing the relationship between the metrics.

The modified indirect metrics, eg, metrics applied on smoothed image data, showed little improvement over the traditional metrics when evaluated using ideal or corrupted segmentations (Fig. 4a,c). In this case the main source of variability was the presence of a mixture of tissues inside the mask, variability that would not be reduced by smoothing. However, substantial improvement was observed when using conservative segmentation, especially in the case of modified CV_{WM} and *CJV* (Fig. 4b). The modified CV_{GM} showed less improvement, probably due to a smaller averaging effect on the thin GM layer, as an isotropic smoothing kernel was used.

Based on the scatterplots, among the traditional metrics the three best performing were CV_{WM} and CV_{GM}, evaluated using conservatively segmented images, and *CJV*, evaluated using ideally segmented images. Among the modified metrics, the three best-performing measures were CV_{WM}, CV_{WM}, and *CJV*, evaluated using conservatively segmented images.

Quantitative Results

Table 2 shows the Spearman correlation coefficients for all noise levels, segmentation types, and metrics. Overall, there was good agreement with the qualitative observations. Conservative segmentation improved the validity of CV_{WM} and CV_{GM}, but *CJV* was best with ideally segmented images. Smoothing had little effect on

Table 2
Spearman Correlation Coefficients Between Direct and Indirect Metrics

		Measures	Noise (%)			Average	
			1	3	5		
Ideal segmentation	Traditional	CV_{WM}	0.46	0.39	0.43	0.43	0.66
		CV_{GM}	0.67	0.77	0.78	0.74	
		CJV	**0.89**	**0.89**	0.65	0.81	
	Modified	CV_{WM}	0.47	0.43	0.39	0.41	0.67
		CV_{GM}	0.74	0.84	**0.86**	**0.83**	
		CJV	0.66	0.79	**0.85**	0.76	
Conservative segmentation	Traditional	CV_{WM}	**0.94**	**0.84**	0.71	**0.83**	0.77
		CV_{GM}	**0.89**	**0.91**	**0.89**	**0.88**	
		CJV	**0.95**	0.66	0.13	0.60	
	Modified	CV_{WM}	**0.96**	**0.97**	**0.96**	**0.96**	**0.94**
		CV_{GM}	**0.87**	**0.90**	**0.90**	**0.89**	
		CJV	**0.97**	**0.97**	**0.95**	**0.96**	
Corrupted segmentation	Traditional	CV_{WM}	0.48	0.40	0.53	0.47	0.51
		CV_{GM}	0.75	0.63	0.52	0.63	
		CJV	0.66	0.45	0.17	0.43	
	Modified	CV_{WM}	0.49	0.44	0.58	0.50	0.71
		CV_{GM}	0.82	0.74	0.75	0.77	
		CJV	**0.87**	**0.86**	0.83	**0.85**	

Values highlighted in bold correspond to correlation of 0.85 and higher. Bold underlined values correspond to correlations exceeding 0.95.

Table 3
Optimization and Comparison of Two Bias Field Correction Algorithms on a BrainWeb MR Volume With 3% Noise and 20% Bias Field Using Ideal Segmentation and Traditional Indirect Performance Measures

Parameter value	Direct	Indirect		
	L_2	CV_{WM}	CV_{GM}	*CJV*
		Histogram sharpening		
30	0.0224	**0.0499**	**0.0993**	0.5325
40	0.0173	0.0505	0.0997	0.5234
50	0.0144	0.0506	0.0994	0.5201
60	0.0123	0.0516	0.0997	**0.5180**
70	0.0124	0.0528	0.1001	0.5184
80	**0.0114**	0.0530	0.1000	0.5181
90	0.0137	0.0542	0.1006	0.5202
100	0.0139	0.0545	0.1007	0.5207
110	0.0168	0.0558	0.1014	0.5240
120	0.0175	0.0562	0.1016	0.5250
130	0.0183	0.0566	0.1018	0.5261
140	0.0192	0.0570	0.1020	0.5273
150	0.0202	0.0574	0.1023	0.5287
160	0.0211	0.0578	0.1026	0.5301
170	0.0220	0.0582	0.1029	0.5315
180	0.0228	0.0585	0.1031	0.5326
190	0.0236	0.0589	0.1034	0.5338
200	0.0244	0.0592	0.1036	0.5349
		Surface fitting		
5	0.0246	0.0571	0.1035	0.5380
10	0.0047	0.0519	0.0991	0.5164
15	0.0036	0.0520	**0.0990**	0.5154
20	0.0034	0.0519	**0.0990**	0.5153
25	**0.0031**	0.0518	0.0991	0.5150
30	0.0034	0.0517	0.0991	0.5150
35	0.0034	0.0516	0.0992	0.5144
40	0.0064	0.0515	0.0998	**0.5135**
45	0.0104	0.0516	0.1007	0.5142
50	0.0163	0.0515	0.1021	0.5194
55	0.0169	0.0515	0.1022	0.5198
60	0.0164	**0.0514**	0.1021	0.5194

Table 4
Optimization and Comparison of Two Bias Field Correction Algorithms on BrainWeb MR Volume With 3% Noise and 20% Bias Field Using Conservative Segmentation and Both Traditional and Modified Indirect Performance Measures

Parameter value	Direct	Indirect					
	L_2	Traditional			Modified		
		CV_{WM}	CV_{GM}	CJV	CV_{WM}	CV_{GM}	CJV
		Histogram sharpening					
30	0.0224	0.0348	0.0566	0.3114	0.0199	0.0353	0.1867
40	0.0173	0.0325	0.0555	0.2880	0.0156	0.0335	0.1580
50	0.0144	0.0316	0.0544	0.2801	0.0136	0.0318	0.1447
60	0.0123	0.0307	0.0538	0.2686	0.0116	0.0307	0.1305
70	0.0124	0.0305	0.0539	0.2632	0.0111	0.0307	0.1261
80	**0.0114**	**0.0304**	**0.0534**	0.2612	**0.0107**	**0.0301**	**0.1229**
90	0.0137	0.0309	0.0541	0.2607	0.0120	0.0310	0.1281
100	0.0139	0.0309	0.0541	**0.2603**	0.0123	0.0311	0.1290
110	0.0168	0.0318	0.0551	0.2626	0.0143	0.0328	0.1392
120	0.0175	0.0320	0.0553	0.2631	0.0148	0.0332	0.1418
130	0.0183	0.0324	0.0556	0.2643	0.0155	0.0337	0.1453
140	0.0192	0.0327	0.0559	0.2655	0.0162	0.0344	0.1488
150	0.0202	0.0330	0.0563	0.2667	0.0169	0.0351	0.1526
160	0.0211	0.0334	0.0568	0.2682	0.0176	0.0357	0.1562
170	0.0220	0.0337	0.0572	0.2697	0.0182	0.0363	0.1594
180	0.0228	0.0340	0.0576	0.2710	0.0187	0.0369	0.1624
190	0.0236	0.0343	0.0581	0.2724	0.0192	0.0376	0.1654
200	0.0244	0.0345	0.0586	0.2738	0.0197	0.0383	0.1684
		Surface fitting					
5	0.0246	0.0330	0.0591	0.2764	0.0167	0.0391	0.1639
10	0.0047	0.0293	0.0532	0.2647	0.0069	0.0296	0.1109
15	0.0036	**0.0292**	**0.0526**	**0.2609**	**0.0068**	**0.0285**	0.1067
20	0.0034	**0.0292**	0.0527	0.2622	**0.0068**	0.0287	0.1077
25	**0.0031**	**0.0292**	0.0527	0.2617	**0.0068**	0.0287	**0.1073**
30	0.0034	**0.0292**	0.0529	0.2632	**0.0068**	0.0290	0.1088
35	0.0034	0.0293	0.0529	0.2630	0.0069	0.0291	0.1091
40	0.0064	0.0294	0.0538	0.2649	0.0075	0.0306	0.1152
45	0.0104	0.0297	0.0552	0.2686	0.0086	0.0331	0.1255
50	0.0163	0.0302	0.0586	0.2831	0.0102	0.0385	0.1485
55	0.0169	0.0305	0.0588	0.2844	0.0109	0.0388	0.1520
60	0.0164	0.0305	0.0584	0.2841	0.0111	0.0382	0.1513

validity for measures evaluated using ideal and corrupted segmentations, but led to substantial improvement in combination with conservatively segmented images. The improvements were particularly notable for images with a large amount of noise (5%). For example, the modified *CJV* had practically the same correlation as the traditional *CJV* (0.97 vs. 0.95) at noise level of 1%; at 5% noise level the smoothing improved correlation from 0.12 to 0.95 (Table 2). The observed improvement at high noise levels is consistent with the prediction that noise reduction makes it easier to detect smaller changes in intensity nonuniformity.

In sum, the modified versions of CV_{WM} and *CJV* evaluated using conservatively segmented images achieved the highest validity across all tested indirect metrics, with Spearman correlation coefficient exceeding 0.95 at all noise levels.

Practical Applications

To illustrate a practical ramification of our findings, consider the results of using traditional and modified indirect measures to assess a subset of corrected synthetic brain images that contained 3% noise and 20% intensity inhomogeneity (Tables 3, 4). Critically, using different traditional validity measures led to widely divergent suggestions regarding the optimal histogram sharpening and surface fitting parameters. On the basis of the direct measure alone, the optimal parameter was 80 for histogram sharpening and 25 for surface fitting. Overall, surface fitting achieved better performance in this example (L_2-norm value of 0.003 for surface fitting vs. 0.011 for the histogram sharpening). Using the traditional indirect measures on ideally segmented images led to completely different conclusions (Table 3). On the basis of CV_{WM} the best parameter values were 30 (histogram sharpening) and 60 (surface fitting) and the former approach outperformed the latter, contradicting the results obtained using the direct measure. Using CV_{GM} and *CJV* correctly identified the best-performing method but led to wrong parameter choices. These findings were consistent with the results in Table 2, which suggested that *CJV* should have been the best-performing metric under these conditions, with CV_{GM} a close second.

When the traditional indirect metrics were evaluated using conservatively segmented images, CV_{WM} and CV_{GM} were able to correctly identify the best-performing method (Table 4) and were better than *CJV* at finding

the optimal parameter values. This is consistent with the finding that *CJV* achieved only a 0.66 correlation with the direct measure (Table 2). Among the modified measures, all three correctly identified the best-performing method. In addition, the modified CV_{WM} and *CJV* produced veridical parameter values for both correction approaches.

DISCUSSION

We have shown that existing indirect measures that assess performance of bias field correction approaches have mediocre validity (correlation of 0.67 on average), which explains why they often lead to conflicting statements regarding a best-performing method. Image noise and inclusion of partial volume voxels and neighboring tissues were implicated as likely reasons. We demonstrated that the combined effect of conservative segmentation and smoothing significantly improves validity.

It is often argued that *CJV* is preferable to CV_{WM} and CV_{GM}. For example, *CV* can be difficult to interpret when it improves for one tissue class but not for others (5,14). It is also possible for a correction method to transform a given image, so that *CV* of two tissues is improved while the overlap between their intensity distributions is increased, making subsequent segmentation difficult (5,15). Our experimental results (Table 2, Figs. 3, 4) suggest that at least for the chosen algorithms and parameter ranges, CV_{WM} is equal and even preferable to *CJV*. We found that *CJV* performed better compared to other metrics when these were evaluated using ideally segmented images. However, the situation was reversed when conservatively segmented images were used, especially when the image noise was large. Even though combining conservative segmentation and smoothing equalizes the quality of CV_{WM} and *CJV*, the former would still be preferable. GM segmentation, necessary for evaluation of *CJV*, is susceptible to errors, whether obtained by an expert or through an automatic segmentation algorithm, diminishing the validity of *CJV* to the level of poorly segmented images.

However, when relying on CV_{WM} alone, a sufficient amount of smoothing should be applied to avoid reduction of contrast between GM and WM, undetectable by CV_{WM}. To achieve this, smoothing kernel size for histogram sharpening should be maintained above 30 mm and polynomial order for surface fitting should be kept below 4.

The results presented here were obtained using simulated data and it is unclear whether they hold for real MR data. For example, human subject MR brain data exhibits regional variation in WM and GM tissue intensity (26), resulting in nonuniformities that cannot be adequately modeled (1). Further research is needed to address this issue.

In conclusion, our findings suggest that assessing the quality of nonuniformity correction from indirect performance measures applied in a traditional fashion can yield inappropriate guidance for parameter selection during inhomogeneity correction. However, if these metrics are used on lightly smoothed images with conservatively segmented tissues, their validity improves considerably, potentially resulting in more appropriate correction of intensity inhomogeneities that could ultimately result in more accurate segmentation of brain tissues.

REFERENCES

1. Belaroussi B, Milles J, Carme S, Zhu YM, Benoit-Cattin H. Intensity non-uniformity correction in MRI: existing methods and their validation. Med Image Anal 2006;10:234–246.
2. Bernstein MA, Huston J, Ward HA. Imaging artifacts at 3.0T J Magn Reson Imaging 2006;24:735–746.
3. Hou Z, Huang S, Hu Q, Nowinski WL. A fast and automatic method to correct intensity inhomogeneity in MR brain images. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Copenhagen, Denmark; 2006.
4. Styner M, Brechbuhler C, Szekely G, Gerig G. Parametric estimate of intensity inhomogeneities applied to MRI. IEEE Trans Med Imaging 2000;19:153–165.
5. Vovk U, Pernus F, Likar B. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans Med Imaging 2007;26: 405–421.
6. Arnold JB, Liow JS, Schaper KA, et al. Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. Neuroimage 2001;13:931–943.
7. Brinkmann BH, Manduca A, Robb RA. Optimized homomorphic unsharp masking for MR grayscale inhomogeneity correction. IEEE Trans Med Imaging 1998;17:161–171.
8. Prima S, Ayache N, Barrick T, Roberts N. Maximum likelihood estimation of the bias field in MR brain image: investigating different modelings of the imaging process. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Utrecht, The Netherlands; 2001.
9. Studholme C, Cardenas V, Song E, Ezekiel F, Maudsley A, Weiner M. Accurate template-based correction of brain MRI intensity distortion with application to dementia and aging. IEEE Trans Med Imaging 2004;23:99–110.
10. Luo J, Zhu Y, Clarysse P, Magnin I. Correction of bias field in MR images using singularity function analysis. IEEE Trans Med Imaging 2005;24:1067–1085.
11. Lewis EB, Fox NC. Correction of differential intensity inhomogeneity in longitudinal MR images. Neuroimage 2004;23:75–83.
12. Boyes RG, Gunter JL, Frost C, et al. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. Neuroimage 2008;39:1752–1762.
13. Dawant BM, Zijdenbos AP, Margolin RA. Correction of intensity variations in MR images for computer-aided tissue classification. IEEE Trans Med Imaging 1993;12:770–781.
14. Gispert JD, Reig S, Pascau J, Vaquero JJ, Garcia-Barreno P, Desco M. Method for bias field correction of brain T1-weighted magnetic resonance images minimizing segmentation error. Hum Brain Mapp 2004;22:133–144.
15. Likar B, Viergever MA, Pernus F. Retrospective correction of MR intensity inhomogeneity by information minimization. IEEE Trans Med Imaging 2001;20:1398–1410.
16. Meyer CR, Bland PH, Pipe J. Retrospective correction of intensity inhomogeneities in MRI. IEEE Trans Med Imaging 1995;14:36–41.
17. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 1998;17:87–97.
18. Van Leemput K, Maes F, Vandermeulen D, Suetens P. Automated model-based bias field correction of MR images of the brain. IEEE Trans Med Imaging 1999;18:885–896.
19. Bansal R, Staib LH, Peterson BS. Correcting nonuniformities in MRI intensities using entropy minimization based on an elastic model. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Saint-Malo, France; 2004.
20. Liew AW, Yan H. An adaptive spatial fuzzy clustering algorithm for 3-D MR image segmentation. IEEE Trans Med Imaging 2003;22: 1063–1075.

21. Wells WM, Grimson WL, Kikinis R, Jolesz FA. Adaptive segmentation of MRI data. IEEE Trans Med Imaging 1996;15:429–442.
22. Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. Neuroimage 2001;13:856–876.
23. Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005; 26:839–851.
24. Johnston B, Atkins MS, Mackiewich B, Anderson M. Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. IEEE Trans Med Imaging 1996;15:154–169.
25. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage 2000;11:805–821.
26. van Walderveen MA, van Schijndel RA, Pouwels PJ, Polman CH, Barkhof F. Multislice T1 relaxation time measurements in the brain using IR-EPI: reproducibility, normal values, and histogram analysis in patients with multiple sclerosis. J Magn Reson Imaging 2003;18:656–664.
27. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15:72–101.
28. Manjon JV, Lull JJ, Carbonell-Caballero J, Garcia-Marti G, Marti-Bonmati L, Robles M. A nonparametric MRI inhomogeneity correction method. Med Image Anal 2007;11:336–345.
29. Cocosco CA, Zijdenbos AP, Evans AC. A fully automatic and robust brain MRI tissue classification method. Med Image Anal 2003;7: 513–527.
30. Collins DL, Zijdenbos AP, Kollokian V, et al. Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging 1998;17:463–468.
31. Kwan RK, Evans AC, Pike GB. MRI simulation-based evaluation of image-processing and classification methods. IEEE Trans Med Imaging 1999;18:1085–1097.
32. Zhuge Y, Udupa JK, Liu J, Saha PK, Iwanage T. Scale-based method for correcting background intensity variation in acquired images. SPIE Medical Imaging: Image Processing; 2002.