Empirical Evaluation of Bias Field Correction Algorithms for Computer-Aided Detection of Prostate Cancer on T2w MRI

Satish Viswanath\(^a\), Daniel Palumbo\(^a\), Jonathan Chappelow\(^a\), Pratik Patel\(^a\), B. Nicolas Bloch\(^b\), Neil Rofsky\(^c\), Robert Lenkinski\(^d\), Elisabeth Genega\(^d\), Anant Madabhushi\(^a\)

\(^a\)Rutgers, the State University of New Jersey, USA; \(^b\)Boston Medical Center, Boston, MA, USA; 
\(^c\)UT Southwestern Medical Center, Dallas, Texas, USA; \(^d\)Beth Israel Deaconess Medical Center, Boston, MA, USA.

ABSTRACT

In magnetic resonance imaging (MRI), intensity inhomogeneity refers to an acquisition artifact which introduces a non-linear variation in the signal intensities within the image. Intensity inhomogeneity is known to significantly affect computerized analysis of MRI data (such as automated segmentation or classification procedures), hence requiring the application of bias field correction (BFC) algorithms to account for this artifact. Quantitative evaluation of BFC schemes is typically performed using generalized intensity-based measures (percent coefficient of variation, \(\%CV\)) or information-theoretic measures (entropy). While some investigators have previously empirically compared BFC schemes in the context of different domains (using changes in \(\%CV\) and entropy to quantify improvements), no consensus has emerged as to the best BFC scheme for any given application. The motivation for this work is that the choice of a BFC scheme for a given application should be dictated by application-specific measures rather than \textit{ad hoc} measures such as entropy and \(\%CV\). In this paper, we have attempted to address the problem of determining an optimal BFC algorithm in the context of a computer-aided diagnosis (CAD) scheme for prostate cancer (CaP) detection from T2-weighted (T2w) MRI. One goal of this work is to identify a BFC algorithm that will maximize the CaP classification accuracy (measured in terms of the area under the ROC curve or AUC). A secondary aim of our work is to determine whether measures such as \(\%CV\) and entropy are correlated with a classifier-based objective measure (AUC). Determining the presence or absence of these correlations is important to understand whether domain independent BFC performance measures such as \(\%CV\), entropy should be used to identify the optimal BFC scheme for any given application. In order to answer these questions, we quantitatively compared 3 different popular BFC algorithms on a cohort of 10 clinical 3 Tesla prostate T2w MRI datasets (comprising 39 2D MRI slices): N3, PABIC, and the method of Cohen et al. Results of BFC via each of the algorithms was evaluated in terms of \(\%CV\), entropy, as well as classifier AUC for CaP detection from T2w MRI. The CaP classifier was trained and evaluated on a per-pixel basis using annotations of CaP obtained via registration of T2w MRI and \textit{ex vivo} whole-mount histology sections. Our results revealed that different BFC schemes resulted in a maximization of different performance measures, that is, the BFC scheme identified by minimization of \(\%CV\) and entropy was not the one that maximized AUC as well. Moreover, existing BFC evaluation measures (\(\%CV\), entropy) did not correlate with AUC (application-based evaluation), but did correlate with each other, suggesting that domain-specific performance measures should be considered in making a decision regarding choice of appropriate BFC scheme. Our results also revealed that N3 provided the best correction of bias field artifacts in prostate MRI data, when the goal was to identify prostate cancer.

Keywords: intensity inhomogeneity, bias field correction, prostate cancer, classification, T2w MRI

1. INTRODUCTION

The most significant acquisition artifact associated with \textit{in vivo} Magnetic Resonance Imaging (MRI) is the smooth variation of signal intensity across the image. This is known as \textit{intensity inhomogeneity}, and refers to the bias field introduced in the image by poor radio frequency coil uniformity, eddy currents, or patient anatomy.\(^1\) Bias field effects in MRI data have been shown to significantly affect computerized quantitative image analysis, e.g. segmentation,\(^2,3\) tissue property measurements,\(^4,5\) as well as automated classification of tissue regions.\(^6\)

A number of post-processing algorithms have been proposed to perform bias field correction (BFC)\(^7-10\) (comprehensively reviewed in [11]). Further, both qualitative and quantitative measures have been proposed in order to determine improvements (if any) offered by BFC for MRI data. Qualitatively, images are visually
evaluated pre- and post-correction to determine improved contrast and resolution of fine anatomic and structural
detail within the image. Qualitative evaluation is, however, largely observer-dependent and tends to be subjective
in nature. Alternatively, BFC algorithms have been evaluated in the context of phantom datasets to which a
known bias field has been applied. The objective in such an experiment is to recover the original data by first
estimating and subsequently correcting for the bias field. One can then precisely and quantitatively calculate
the associated error. In the context of clinical data, the issue of validating BFC in this manner is difficult since
the true amount of intensity inhomogeneity is unknown, making it nearly impossible to precisely determine the
accuracy of correction.

Performance measures that have been previously employed for quantitatively evaluating BFC algorithms have
included (1) percent coefficient of variation (%CV), an image intensity-based statistic to determine whether
the smooth variation in image intensities introduced by the bias field has been reduced after correction, and (2)
entropy, a generalized information theoretic measure which should be lowered after correction. The objective
of these performance measures (%CV, entropy) is to evaluate the performance of BFC schemes in a manner
independent of the domain or application. In certain cases, more domain-specific performance measures have
been employed for BFC evaluation. For instance, in [5], Velthuizen et al evaluated segmentation accuracy pre-
and post-BFC. Similarly in [12], Zijdenbos et al evaluated the similarity index, a measure of tissue self-similarity
post-BFC. Work done in [5,12] suggests that the choice of a BFC scheme should be directed by the overarching
objective or goal of the domain-specific task or application being considered.

Velthuizen et al compared 4 different BFC algorithms in the context of tumor volume response measurements
(via segmentation) in brain MRI data. The algorithms showed significant differences in terms of the bias fields
estimated, but did not show significant differences in terms of the final tumor segmentation (and hence volume
response measurements). The authors concluded that BFC did not necessarily show significant beneficial effects
for their specific application. In [13], 6 different BFC algorithms were compared by visually examining corre-
sponding results as well as quantifying their performance on synthetic phantom datasets (as detailed previously).
None of the algorithms were found to perform ideally under all circumstances. However both studies suggested
that BFC techniques as well as evaluation of these methods required more understanding and study.

In this work, we will explore (a) the role of BFC, and (b) performance measures for evaluating and selecting
the optimal BFC scheme in the context of T2-weighted (T2w), endo-rectal prostate MRI, where the overall goal
is to develop a computerized image-based classifier for accurate prostate cancer (CaP) detection in vivo. In
the clinic, visual analysis of prostate T2w MRI (to determine CaP presence and extent) is done by manually
applying appropriate windowing operations to correct for signal inhomogeneity. However, CaP detection on T2w
MRI is complicated by the presence of benign confounders including prostatitis, fibro-muscular benign prostatic
hyperplasia, or post-biopsy hemorrhagic change. The presence of a bias field further exacerbates the problem
of resolving benign confounders from true CaP regions on T2w endo-rectal prostate MRI due to peripheral
regions of the gland (regions close to the endo-rectal probe) getting masked by a high signal “flare”, as well
as a marked signal intensity decrease in regions farther away from the probe. Figures 1(b) and (d), showing
intensity distributions of cancer and non-cancer regions on T2w endo-rectal prostate MRI before and after BFC
(Figures 1(a) and (c)), clearly reveal that the separation between the CaP and non-CaP classes is significantly
enhanced post-BFC. This also suggests that a computer-aided diagnosis (CAD) scheme for CaP detection on
T2w prostate MRI would benefit from prior application of BFC.

While some groups have recently begun to develop CAD schemes for CaP detection from multi-parametric
MRI, few have explicitly attempted to address the BFC issue for prostate T2w MRI. Vos et al attempted to use T2w and proton density images in conjunction with a known sequence model to correct for the
signal inhomogeneity within the T2w prostate MR image. Other groups have made use of quantitative T2
maps (rather than T2w MRI directly). These approaches result in MR images of lower resolution and require
an additional acquisition for BFC.

Based on the above, it is clear that BFC techniques may be of significant utility in accounting for intensity
inhomogeneity in prostate T2w MRI, which in turn would allow for (1) improved accuracy in discriminating
between tissue classes (CaP and benign regions) in the data, and (2) circumvent the need for complex acquisition
procedures in order to correct artifacts in such data. However, given the large number of BFC algorithms which
Figure 1. Endo-rectal, in vivo T2w prostate MR image (gland has been manually segmented) (a) prior to and (c) post-BFC, with CaP region outlined in red. Intensity distributions for CaP and benign regions superposed on the same axis (b) prior to and (d) post-BFC. Note that the intensity distributions for the 2 classes are further apart post-BFC ((d)) compared to pre-BFC ((b)).

As previously proposed,\textsuperscript{11} it is not immediately clear how to determine which algorithm provides the “best” bias field correction for prostate T2w MR imagery.

As previously discussed, the determination of the “best” BFC method is largely based on the choice of evaluation measure used. Domain-independent evaluation measures (%CV and entropy) attempt to measure changes in the image intensities post-correction. For the problem considered in this work, where the objective is to maximize the area under the ROC curve (AUC) post-correction, it is not immediately clear whether a BFC algorithm that is selected based on maximally minimizing %CV and entropy would be the appropriate choice. The objectives of this paper are thus three-fold:

(a) What is the best bias field correction method for T2w MR imagery, given a specific CAD related task (e.g. CaP detection on T2w MRI)?

(b) Do existing BFC evaluation measures (%CV, entropy) accurately reflect CAD performance? Are application-based BFC performance measures (AUC) independent and uncorrelated with respect to existing measures?

(c) Would all BFC evaluation measures select the same BFC algorithm as most optimal? In the case of conflicts between different evaluation measures, how should one pick the appropriate BFC method?

With the goal of answering these questions, in this work we shall compare the effect of 3 popular BFC algorithms: (1) N3,\textsuperscript{8} (2) the low-pass filtering method proposed by Cohen et al,\textsuperscript{7} and (3) PABIC,\textsuperscript{9} on several 3 Tesla (T) in vivo endo-rectal prostate T2w MRI studies. Each BFC algorithm was evaluated in the context of our CAD application (via classifier AUC) as well as via %CV and entropy, pre- and post-BFC, in order to identify the best overall BFC scheme. Additionally, a correlation analysis was performed across the 3 different performance measures to evaluate the most uncorrelated quantities.

The rest of paper is organized as follows. Section 2 describes each of the BFC algorithms considered in this study, while Section 3 describes the BFC performance measures. Section 4 describes the specific CAD application (CaP detection from T2w MRI) considered in this work, while Sections 5 and 6 showcase the results of the experiments conducted and concluding remarks.

2. OVERVIEW OF THE ISSUE OF BIAS FIELD AND ASSOCIATED CORRECTION TECHNIQUES

As previously described, intensity inhomogeneity describes the problem in MR imagery where the signal intensities in the image vary smoothly, leading to obfuscation of anatomical detail.\textsuperscript{1} A common reason for presence of such image intensity variations is due to a corresponding variation in the RF excitation field, applied to the image during MR acquisition.\textsuperscript{5} For an MR scene \( C = (C, f) \), where \( f(c) \) is the measured signal intensity associated with every voxel \( c \) in a 3D grid \( C \), intensity inhomogeneity may be modeled as follows,\textsuperscript{7–9}

\[
f(c) = \beta(c)\tilde{f}(c) + \eta(c),
\]  

(1)
where \( \tilde{f}(c) \) is the true, unknown MR signal intensity associated with location \( c \in C \), \( \beta(c) \) is the unknown multiplicative, smoothly varying bias field, and \( \eta(c) \) is the additive white Gaussian noise assumed to be independent of \( \tilde{f}(c) \). If one assumes that the signal term is much greater than the noise term, the above equation (in log-space) reduces to

\[
\log\tilde{f}(c) \approx \log f(c) - \log\beta(c).
\]

Note that calculations in log space allow us to perform the correction as a simple subtraction followed by exponentiation in order to recover an approximation of the true signal intensity, denoted \( \hat{f}(c) \). We now examine the different BFC techniques considered in this work.

### 2.1 N3

The N3 technique\(^*\) is a non-parametric nonuniform intensity normalization scheme,\(^8\) which does not require extended scan time or expert supervision. We denote the associated distributions of \( f(c) \), \( \beta(c) \), and \( \tilde{f}(c) \) as \( F \) (measured distribution of signal intensities), \( B \) (smoothly varying bias field), and \( \tilde{F} \) (unknown true distribution of signal intensities), respectively. N3 assumes that the bias field \( B \) blurs the true intensity distribution \( \tilde{F} \), resulting in \( F \). N3 attempts to recover the un-blurred version of \( \tilde{F} \) via (1) iteratively calculating an estimate of \( \tilde{F} \) by sharpening \( F \), and (2) incrementally de-convolving smooth field estimates \( B \) from the measured \( F \) in order to best approximate the estimated \( \tilde{F} \). This is continued until no change occurs in the estimates of \( B \) and \( \tilde{F} \).

### 2.2 Low-pass filtering (LPF)

In [7], Cohen et al presented an implementation of the low-pass filtering approach first examined in [21]. The underlying assumption for this approach is that the bias field \( \beta(c) \) may be estimated via convolution of a smoothing Gaussian kernel with the image, i.e. a low-pass filtering of the signal. This allows for reformulation of Equation 2 as,

\[
\hat{f}(c) = \exp\{\log[f(c)] - \text{lpf}(\log[f(c)])\},
\]

where \text{lpf} is a low-pass filtering operation.

Prior to correction, the acquired image undergoes pre-processing to eliminate the background as well as to remove extreme intensity values. \text{lpf}(\log[f(c)]) is then calculated and provides a rough estimate of the bias field. This estimated bias field is then removed from the acquired image, while normalizing signal intensities such that the average pixel intensity (in the image) is retained after correction.

### 2.3 Parametric Bias Field Correction (PABIC)

The parametric bias field correction method presented by Styner et al\(^9\) has been implemented as the ITK BiasCorrector algorithm\(^1\). Here, a parametric model of tissue class statistics is utilized with the underlying assumption that every pixel in the image is assigned to a class whose statistics are known \textit{a priori}. A prerequisite of the method is hence the knowledge of existing classes in the data and their corresponding statistics.

The multiplicative bias field \( \beta(c) \) is approximated via a linear combination of smooth Legendre polynomials. An evolution strategy-based optimization is used to determine the order and parameters of these Legendre polynomials in order to correct for the bias field; the aim of PABIC being to assign pixels image intensity values which closely mirror their corresponding class statistics (such as class means).

### 3. EVALUATING THE EFFECTS OF BIAS FIELD CORRECTION

#### 3.1 Statistical measures

The most popular\(^1,8,22\) measure for evaluating the effectiveness of BFC procedures is the percent coefficient of variation, defined as,

\[
\phi^{CV} = \frac{\sigma_t}{\mu_t} \times 100,
\]

where \( \sigma_t \) and \( \mu_t \) are the mean and standard deviation of image intensities within the region of interest \( t \). Madabhushi et al\(^22\) and others\(^8\) have shown that \( \phi^{CV} \) typically tends to reduce within \( t \) after bias field correction.

\(^*\)http://www.nitrc.org/projects/nu_correct/

\(^1\)http://www.itk.org/ITK/applications/MRIBiasCorrection.html
3.2 Information theoretic measures
The underlying assumption for entropy as a BFC evaluation measure is that the intensity non-uniformity introduced by the bias field will raise the entropy of the data distribution.\(^1\) Entropy is given by

\[
\phi^{\text{Ent}} = \sum_{i=1}^{G} \left[P(g_i) \times \log[P(g_i)]\right],
\]

where \(G\) is the total number of unique graylevel intensities \(g_i\) in the image, each of which have an associated probability of occurrence, \(P(g_i)\). Ideally, a BFC algorithm should reduce \(\phi^{\text{Ent}}\) post-correction.

3.3 Classifier-specific measures
As described in the introduction, if the objective of BFC is to improve classifier accuracy, then it should follow that a classifier-based performance measure (such as area under the ROC curve) should increase following BFC. Consider a binary class problem, where for a given scene \(C = (C, f)\), ground truth class labels are available for every voxel \(c \in C\), \(Y(c) \in \{0, 1\}\). Denoting the target class of pixels as \(G(C) = \{c | Y(c) = 1\}\), the objective of the classifier is to identify the set \(G(C)\) as accurately as possible. If a probabilistic classifier is applied to the scene \(C\), a classifier result \(h(c) \in \{0, 1\}\) will be obtained at every \(c \in C\). \(h(c)\) can be thresholded at different values \(\rho \in [0, 1]\) to obtain the hard classification result \(h^\rho(c), c \in C\). If \(|h^\rho(c) - Y(c)| = 0\), then \(h^\rho\) can be said to have assigned the correct class label to \(c \in C\), otherwise not.

An ROC curve which represents the trade-off between classification sensitivity and specificity can now be generated for \(h\) (based on \(h^\rho\)). The vertical axis of the ROC curve is the true positive rate (or sensitivity, \(SP\)), and the horizontal axis is the false positive rate (or 1-specificity, \(SP\)). We denote the set of hard classification results (at every \(\rho\)) for the entire scene \(C\) as \(\Psi^\rho = \{c | h(c) \geq \rho\}, \rho \in [0, 1]\). Each point on the ROC curve will correspond to the \(SN\) and \(SP\) of \(\Psi^\rho\) at some \(\rho \in [0, 1]\). Given the target class \(G(C)\), for each \(\Psi^\rho, SN^\rho\) and \(SP^\rho\) are calculated as

\[
SN^\rho = 1 - \frac{|G(C) - \Psi^\rho|}{|G(C)|} \quad \text{and} \quad SP^\rho = 1 - \frac{|\Psi^\rho - G(C)|}{|C - G(C)|}.
\]

The area under the ROC curve (AUC, denoted via \(\phi^{\text{AUC}}\)) can then be calculated, with a value of \(\phi^{AUC} = 1\) representing perfect classification performance.

4. CAD APPLICATION: DETECTING PROSTATE CANCER FROM T2-WEIGHTED MRI

In this section we describe the specifics of our CAD application, previously presented in [19, 20], which will be employed to evaluate the results of different BFC algorithms.

4.1 Data Description and Notation
A total of 10 in vivo pre-operative endo-rectal T2w MRI patient datasets were acquired from the Beth Israel Deaconess Medical Center using a 3 Tesla Genesis Signa MRI machine. These patients had previously been diagnosed with prostate cancer via core needle biopsies. After a radical prostatectomy procedure, the whole mount histology specimens (WMHS) were sectioned and stained with Haematoxylin and Eosin (H & E) so that a pathologist could delineate presence and extent of CaP. 37 MRI-WMHS pairs were then identified via our recently developed group-wise multi-modal slice correspondence matching scheme.\(^2\) These slice correspondences were later confirmed by a pathologist and radiologist working together. Our previously presented automated non-linear registration scheme, COLLINARUS,\(^3\) was used to align corresponding T2w MRI and WMHS data in order to map pathologist-annotated CaP regions onto corresponding MRI slices.

We denote an uncorrected prostate T2w MR image as \(C = (C, f)\), where \(f(c)\) assigns an intensity value to every pixel \(c \in C\). The corrected scenes are denoted as \(C^i = (C, f^i), i \in \{1, 2, 3\}\), where (1) \(C^1\) corresponds to PABIC,\(^4\) (2) \(C^2\) corresponds to N3,\(^5\) and (3) \(C^3\) corresponds to the method of Cohen et al\(^7\) (referred to as LPF). \(G(C)\) denotes the set of pixels labeled as CaP via registration with WMHS, and is considered as the surrogate for ground truth CaP extent on T2w MRI. Hence if \(c \in G(C), Y(c) = 1\); \(Y(c) = 0\) otherwise.
4.2 Feature extraction

A total of 69 texture features previously shown to improve differentiation between CaP and benign regions were extracted from $C, C_i^k, i \in \{1, 2, 3\}$, on a per-pixel basis. These features are obtained by (1) calculating responses to various filter operators, and (2) computing gray level intensity co-occurrence statistics, as follows,

1. **Non-steerable gradient features**: Eight non-steerable gradient features were obtained by convolving Sobel and Kirsch edge filters and first-order spatial derivative operators with every scene $C, C_i^k, i \in \{1, 2, 3\}$. These operators detect the strength of horizontal, vertical, and diagonal edges within the image using linear kernels.

2. **Steerable gradient features**: Gabor operators comprise the steerable class of gradient calculations which attempt to match localized frequency characteristics. A Gabor filter can be defined as the modulation of a complex sinusoid by a Gaussian function and is controlled by scale, bandwidth, and frequency parameters. 48 Gabor features were calculated based on responses to convolving every scene $C, C_i^k, i \in \{1, 2, 3\}$, with distinct Gabor operators obtained by varying each of the scale, bandwidth, and frequency parameters.

3. **Second order statistical features**: Second order statistical features have been proposed by Haralick and have found wide application in computing features with perceptual meaning in CAD systems. These features are based on quantifying the spatial gray-level co-occurrence within local neighborhoods around each pixel in an image, stored in the form of matrices. 13 Haralick features were calculated based on statistics derived from these matrices for each $C, C_i^k, i \in \{1, 2, 3\}$.

We direct the reader to [19, 20] for a more detailed description of the individual texture features. Feature extraction results in feature scenes $F_{i,\varphi} = (C, f_{i,\varphi}), i \in \{1, 2, 3\}$, where $f_{i,\varphi}(c)$ is the feature value at location $c \in C$ when feature operator $\varphi$ is applied to scene $C_i^k$. Based on aggregating feature values, every pixel $c \in C$ is associated with a set of 3 distinct feature vectors, $F_i(c) = [f_{i,\varphi}(c)|\varphi \in \{1,\ldots,69\}], i \in \{1, 2, 3\}$. We also construct the feature vector $F(c)$ by aggregating feature values corresponding to the uncorrected scene $C$.

4.3 Classification

A random forest (RF) classifier to discriminate voxels $c \in C$ as belonging to CaP or benign regions is constructed based off $F(c), F_i^k(c), i \in \{1, 2, 3\}$, and corresponding class label information $Y(c)$. RFs are a classifier ensemble of decision trees based on bootstrap aggregation (or bagging), which combines the results of multiple weak classifiers such that the overall bias and variance across all classifiers is reduced.

Based on classifying the different feature vectors $F(c), F_i^k(c), i \in \{1, 2, 3\}$, associated with every pixel $c \in C$, the associated voxel-level probabilistic RF classifier result is denoted as $h(c), h_i^k(c), i \in \{1, 2, 3\}$, respectively. Thresholding $h(c), h_i^k(c), c \in C$, for different $\rho$ values, we obtain the corresponding classification results $\psi_\rho, \psi_i^k, \rho \in [0, 1], i \in \{1, 2, 3\}$. A leave-one-out, cross-validation approach was adopted for the classifier, where at each iteration a single slice was held out for testing and the remaining were used in training. $\phi^{AUC}$ values were hence calculated based on evaluating the classification of each of 37 slices, and averaging over all 37 cross-validation runs. A final classification performance measure $\phi^{AUC}$ (for the uncorrected scene $C$) and $\phi_i^{AUC}$ (for the corrected scenes $C_i^k), i \in \{1, 2, 3\}$, were thus obtained.

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Experiment 1: Comparing PABIC, N3, and LPF in terms of BFC evaluation measures

In this experiment, $\%CV$, entropy, and AUC were calculated for each of $C, C_i^k, i \in \{1, 2, 3\}$, yielding corresponding values $\phi^\beta, \phi_i^\beta, \beta \in \{CV, Ent, AUC\}$, respectively. This was done for all 37 MR images considered. One-way analysis of variance (ANOVA) tests were performed across (1) $\phi^{CV}, \phi_i^{CV}, i \in \{1, 2, 3\}$, (2) $\phi^{Ent}, \phi_i^{Ent}, i \in \{1, 2, 3\}$, (3) $\phi^{AUC}, \phi_i^{AUC}, i \in \{1, 2, 3\}$, under the null hypothesis that there was no difference between evaluation measures in each comparison. Note that the ANOVA test was intended to determine statistical significance between data pre- and post-BFC, as well as between different BFC results.
Figure 2. (a), (e) Uncorrected prostate T2w MRI image (C) and corresponding bias field corrected images (b), (f) $C_1^\kappa$ (PABIC), (c), (g) $C_2^\kappa$ (N3), and (d), (h) $C_3^\kappa$ (LPF), respectively. Note the improved image intensity contrast and resolvability of structural detail for after BFC ((b)-(d), (f)-(h)) compared to no BFC ((a), (e)).

Figure 2 shows images pre- and post-correction for 2 different endo-rectal prostate T2w MR images. The post-corrected scenes $C_1^\kappa, C_2^\kappa, C_3^\kappa$ (Figures 2(b)-(d), (f)-(h)) appear to have improved image intensity contrast compared to $C$ (Figures 2(a), (e)). Figure 3 shows the CaP probability heatmaps corresponding to 2 prostate images, where the classifier was applied both pre- and post-correction (by each of N3, PABIC, and LPF). The corresponding likelihood maps following BFC (Figures 3(b)-(d), (f)-(h)) appear to delineate CaP areas with

Figure 3. Probability images for CaP extent (via a RF classifier) corresponding to (a), (e) $C$, (b), (f) $C_1^\kappa$ (PABIC), (c), (g) $C_2^\kappa$ (N3), and (d), (h) $C_3^\kappa$ (LPF), respectively. Ground truth CaP extent for these images (obtained by registering the pre-operative T2w MRI with the post-operative WMHS) have been outlined in red on (a)-(h). Note the more specific and accurate classification obtained following N3 ((c), (g)) compared to any either of $C, C_1^\kappa, C_3^\kappa$. 
higher sensitivity and specificity compared to before BFC (Figures 3(a), (e)).

The average \( \phi^{CV} \), \( \phi^{Ent} \), and \( \phi^{AUC} \) (and corresponding standard deviations) for the 3 BFC algorithms with respect to uncorrected data are summarized in Table 1. We see that \( \phi^{CV}_{3} \) is lower compared to \( \phi^{CV} \), though the reduction in the case of \( \phi^{CV}_{3} \) (N3) is less compared to \( \phi^{CV}_{1} \) (PABIC) and \( \phi^{CV}_{2} \) (LPF). Corresponding results for \( \phi^{Ent} \) are similar. ANOVA tests showed statistically significant values \( (p < 0.05) \) between all of \( \{ \phi^{CV}, \phi^{CV}_{i}, \phi^{CV}_{3} \} \), \( i \in \{1, 2, 3\} \), as well as \( \{ \phi^{Ent}, \phi^{Ent}_{i}, \phi^{Ent}_{3} \} \) (Table 1).

While differences in classification performance are observed for all 3 correction algorithms \( \{ \phi^{CV}_{1}, \phi^{CV}_{2}, \phi^{CV}_{3} \} \) compared to the uncorrected data \( \phi^{CV} \), only \( \phi^{AUC}_{2}, \phi^{AUC}_{3} \) show improvements over \( \phi^{AUC} \). PABIC does not show any improvement in \( \phi^{AUC}_{1} \) despite having significantly reduced \( \phi^{CV}_{1} \) and \( \phi^{Ent}_{1} \) values, compared to the uncorrected scenes. In contrast, N3 yields the highest \( \phi^{AUC} \), despite \( \phi^{CV}_{2} \) and \( \phi^{Ent}_{2} \) showing lesser reduction as compared to \( \phi^{CV}_{1}, \phi^{CV}_{3}, \phi^{Ent}_{1}, \phi^{Ent}_{3} \), respectively.

Therefore while \%CV and entropy suggest N3 is not as optimal for BFC compared to PABIC and LPF, \( \phi^{AUC} \) suggests that N3 is optimal from the perspective of our CAD application \( (AUC=0.79) \). Hence while PABIC might have been selected according to \%CV and entropy criteria, the corresponding AUC \( (AUC=0.65) \) is actually lower compared to that for the uncorrected scene \( (AUC=0.72) \). These results suggest that the criterion for selecting the optimal BFC scheme needs to be dictated by the domain-specific application for which BFC is employed.

### 5.2 Experiment 2: Determining correlation between \%CV, entropy, and AUC

The objective of this experiment was to determine whether our application-based evaluation of BFC (via AUC) was independent and uncorrelated from \%CV and entropy, respectively. First, trends in each evaluation measure were calculated with respect to uncorrected data in terms of absolute differences as \( |\phi^{\beta} - \phi^{CV}|, \{1, 2, 3\}, \beta \in \{CV, Ent, AUC\} \). Then, the trends for each \( \beta \) were compared on a per-method basis \( (e.g. \ |\phi^{CV} - \phi^{\beta}| \) vs \( |\phi^{AUC} - \phi^{AUC}|) \). Finally, the coefficient of determination, \( R^{2} \), was calculated for each such comparison. Note that a value of \( R^{2} = 1 \) corresponds to perfect correlation between the 2 variables being compared.

Figure 4 displays correlation plots between (1) AUC vs \%CV, (2) AUC vs entropy, and (3) \%CV vs entropy, compared for PABIC, N3, and LPF. For both PABIC (Figure 4(c)) and LPF (Figure 4(i)), we note that a high correlation exists between \( \phi^{CV} \) and \( \phi^{Ent} \). In contrast, \( \phi^{AUC} \) shows low correlation with either of \( \phi^{CV} \) or \( \phi^{Ent} \), across all 3 BFC algorithms (Figures 4(a)-(b), (d)-(e), (g)-(h)). This suggests that \( \phi^{AUC} \) may be an uncorrelated measure of BFC performance as compared to \%CV and entropy.

### 5.3 Experiment 3: Determining correlation between N3, PABIC, and LPF

The aim of this experiment was to determine whether different BFC methods optimize the same evaluation measures. In other words, whether different BFC evaluation criteria would result in the selection of the same BFC algorithm. Using the absolute differences calculated in Experiment 2, we compared trends for each measure between BFC algorithms \( (e.g. \ |\phi^{CV}_{1} - \phi^{CV}_{2}| \) vs \( |\phi^{CV}_{2} - \phi^{CV}_{3}|) \).

Figure 5 displays correlation trends between PABIC, N3, and LPF, evaluated for each of \( \phi^{CV}, \phi^{Ent}, \phi^{AUC} \). High correlation can be observed between PABIC and LPF for \( \phi^{CV} \) and \( \phi^{Ent} \) (Figures 5(c) and (f)), which implies that these algorithms appear to yield a similar correction result. The 2 methods however have very different trends with respect to \( \phi^{AUC} \) (Figure 5(i)). Similar trends in \( \phi^{AUC} \) are observed for both LPF and N3.

Table 1. \( \phi^{CV}, \phi^{Ent}, \phi^{AUC} \) values averaged over 37 MR images for each of \( C, C'_{i}, i \in \{1, 2, 3\} \). \( p \)-values for the ANOVA tests performed show statistically significant differences \( (p < 0.05) \) between all of \( C, C'_{i}, i \in \{1, 2, 3\} \), in terms of both \( \phi^{CV} \) and \( \phi^{Ent} \) (null hypothesis was that there was no difference between \( C, C'_{i}, i \in \{1, 2, 3\} \) in terms of each measure, over all 37 MR images considered).

<table>
<thead>
<tr>
<th>( C )</th>
<th>( \phi^{CV} )</th>
<th>( \phi^{Ent} )</th>
<th>( \phi^{AUC} )</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_{1} )</td>
<td>55.53±11.82</td>
<td>6.72±0.53</td>
<td>0.72±0.14</td>
<td>5.93E-23</td>
</tr>
<tr>
<td>( C_{2} )</td>
<td>29.15±4.79</td>
<td>6.27±0.49</td>
<td>0.65±0.13</td>
<td>9.75E-44</td>
</tr>
<tr>
<td>( C_{3} )</td>
<td>52.09±11.82</td>
<td>6.67±0.54</td>
<td>0.79±0.16</td>
<td>0.57</td>
</tr>
<tr>
<td>( C_{5} )</td>
<td>26.61±2.74</td>
<td>5.99±0.47</td>
<td>0.79±0.15</td>
<td>0.57</td>
</tr>
</tbody>
</table>

55.53, 0.57, 6.67, 0.79, 6.72, 9.75E-44, 37 MR images considered.

Proc. of SPIE Vol. 7963  79630V-8
Figure 4. Correlation plots between BFC evaluation measures ($\phi_{CV}$, $\phi_{Ent}$, $\phi_{AUC}$) for (a)-(c) PABIC, (d)-(f) N3, and (g)-(i) LPF. High correlation is observed between $\phi_{CV}$ and $\phi_{Ent}$ ($R^2$ close to 1), while $\phi_{AUC}$ appears to be largely uncorrelated with $\phi_{CV}$ and $\phi_{Ent}$, suggesting that $\phi_{AUC}$ is an independent evaluation measure of BFC performance.

(Figure 5(g)), even though trends in $\phi_{CV}$ and $\phi_{Ent}$ for N3 are different compared to LPF or PABIC (Figures 5(a)-(b), (d)-(e), (h)). These results suggest that entropy and %CV may result in selecting the same BFC algorithm as optimal.

6. CONCLUDING REMARKS

While a number of bias field correction (BFC) algorithms have been proposed in the literature, there have been relatively few studies exploring the criterion for identifying the optimal BFC algorithm for a specific domain or
Figure 5. Correlation plots between BFC algorithms (PABIC, N3, LPF) for (a)-(c) $\phi_{CV}$, (d)-(f) $\phi_{Ent}$, and (g)-(i) $\phi_{AUC}$. Note similarity in performance for PABIC and LPF in terms of (c) $\phi_{CV}$ and (f) $\phi_{Ent}$, despite the fact that the corresponding trends in (i) $\phi_{AUC}$ are different.

In this work, we have performed a rigorous qualitative and quantitative evaluation of 3 different BFC algorithms as applied to clinical prostate T2w MRI data, to determine how best to correct such data to ensure optimal automated classification of prostate cancer. The primary objectives of this work were to determine (1) which BFC algorithm best corrects MRI data from the perspective of this CAD application, (2) whether existing BFC evaluation measures (such as %CV, entropy) are correlated with the maximization of the domain-specific objective (in this case maximizing classifier performance), and (3) whether the BFC scheme determined as optimal based on one criterion (e.g. %CV or entropy) will also be optimal in terms of the domain-specific criterion (e.g. maximizing area under the ROC curve). Our experiments in comparing 3 different BFC schemes across 10 T2w prostate MRI studies revealed that while a BFC scheme might appear optimal with respect to
one or more criteria, it was not guaranteed to be optimal with respect to all the evaluation criteria. For instance, while the PABIC scheme resulted in the minimization of both CV and entropy post-correction (as expected), classifier performance (CaP detection on T2w MRI) actually decreased (BFC should ideally result in an increase in classifier performance). This may be on account of the fact that PABIC requires prior knowledge of the number of existing classes in the data and their corresponding statistics; difficult to ascertain in the context of prostate T2w MRI. Interestingly, if the application-specific measure had not been considered for evaluation, PABIC might have been selected as the BFC method of choice (based off the decrease in CV and entropy). This implies that bias field correction should also be evaluated in terms of the domain-specific objective, rather than generalized measures alone.

ACKNOWLEDGMENTS

This work was made possible via grants from the Wallace H. Coulter Foundation, National Cancer Institute (Grant Nos. R01CA136535, R01CA140772, and R03CA143991), Department of Defense Prostate Cancer Research Program (W81XWH-08-1-0072), and The Cancer Institute of New Jersey.

REFERENCES