Task-based evaluation of segmentation algorithms for diffusion-weighted MRI without using a gold standard

2012 Phys. Med. Biol. 57 4425

(http://iopscience.iop.org/0031-9155/57/13/4425)

IOP Publishing Physics in Medicine and Biology

Phys. Med. Biol. **57** (2012) 4425–4446 doi:10.1088/0031-9155/57/13/4425

Task-based evaluation of segmentation algorithms for diffusion-weighted MRI without using a gold standard

Abhinav K Jha[1], Matthew A Kupinski[1,2], Jeffrey J Rodríguez[3], Renu M Stephen[4] and Alison T Stopeck[4]

[1] College of Optical Sciences, University of Arizona, Tucson, AZ, USA
[2] Department of Radiology, University of Arizona, Tucson, AZ, USA
[3] Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA
[4] Arizona Cancer Center, University of Arizona, Tucson, AZ, USA

E-mail: akjha@email.arizona.edu

Received 24 February 2012, in final form 4 May 2012
Published 20 June 2012
Online at stacks.iop.org/PMB/57/4425

Abstract

In many studies, the estimation of the apparent diffusion coefficient (ADC) of lesions in visceral organs in diffusion-weighted (DW) magnetic resonance images requires an accurate lesion-segmentation algorithm. To evaluate these lesion-segmentation algorithms, region-overlap measures are used currently. However, the end task from the DW images is accurate ADC estimation, and the region-overlap measures do not evaluate the segmentation algorithms on this task. Moreover, these measures rely on the existence of gold-standard segmentation of the lesion, which is typically unavailable. In this paper, we study the problem of task-based evaluation of segmentation algorithms in DW imaging in the absence of a gold standard. We first show that using manual segmentations instead of gold-standard segmentations for this task-based evaluation is unreliable. We then propose a method to compare the segmentation algorithms that does not require gold-standard or manual segmentation results. The no-gold-standard method estimates the bias and the variance of the error between the true ADC values and the ADC values estimated using the automated segmentation algorithm. The method can be used to rank the segmentation algorithms on the basis of both the ensemble mean square error and precision. We also propose consistency checks for this evaluation technique.

(Some figures may appear in colour only in the online journal)

1. Introduction

Diffusion can be described as the thermally induced behavior of molecules moving in a microscopic random pattern in a field. Diffusion-weighted magnetic resonance imaging (DWMRI) is sensitive to this microscopic motion (Bammer 2003, Huisman 2003) and,

therefore, measures the mobility of water in tissues. This mobility is quantified using the apparent diffusion coefficient (ADC) parameter. Generally, cellular structures restrict water movement, so a decrease in lesion size in response to therapy typically results in a change in the ADC value (Chenevert *et al* 2000). In previous studies, the ADC value has been shown to be a positive indicator to tumor response to therapy, both pre-clinically and clinically (Galons *et al* 1999, Bortner and Cidlowski 2002, Theilmann *et al* 2004, Stephen *et al* 2011, Stephen *et al*). However, for DWMRI to function as an imaging biomarker, the ADC must be estimated accurately. Accurate ADC estimation is a challenge due to noise corruption of the diffusion-weighted (DW) image, contamination due to flow artifacts and ghosting and other random phenomena. In visceral organs such as the liver, pancreas and spleen, this task is even more complicated due to movement of the lesion across different scans (Theilmann *et al* 2004, Stephen *et al* 2011, Stephen *et al*). To circumvent the issues due to movement of the lesion, in many studies, the ADC value of the lesion in visceral organs is computed by segmenting the lesion and subsequently determining its mean signal intensity at the different b values (Theilmann *et al* 2004, Stephen *et al* 2011, Stephen *et al,* Muhi *et al* 2009, Jha *et al* 2010a, 2010b). The lesion mean signal intensity is invariant to organ movement. Using this mean signal intensity, the ADC value of the lesion is determined. However, for this scheme to work, the first-required step is accurate lesion segmentation. Segmenting the lesion manually is a time-consuming and error-prone task (Krishnamurthy *et al* 2004, Jha 2009, Jha *et al* 2010d), so automated segmentation algorithms are required. There is ongoing research and published literature on developing automated segmentation techniques for DW images (Jha *et al* 2010d, Krishnamurthy *et al* 2004, Jha 2009, Mohan *et al* 2010, Saad *et al* 2010, Wu and Jie 2003, Hadjiprocopis *et al* 2005). To compare the performance of various automated segmentation algorithms in order to choose the best one, appropriate evaluation methods are required.

In the field of medical image segmentation, characterizing the performance of segmentation algorithms is an important research problem. Evaluation of segmentation algorithms is a complicated task due to multiple reasons (Chalana and Kim 1997). The first reason is the absence of a gold standard; typically, the only standards available for comparison are segmentations produced by expert observers, but even they suffer from observer bias and intra- and inter-expert variability (Warfield *et al* 2004, Klauschen *et al* 2009, Udupa *et al* 2006, Filippi *et al* 1995, Tunariu *et al* 2010, Lu *et al* 2005). Moreover, the precision of manual segmentations depends on the crispness of the boundaries, the window-level settings for image display, the computer monitor and its settings and the operator's vision characteristics (Udupa *et al* 2006). The other difficulty in comparing segmentation algorithms is defining a metric to compare the computer-generated segmentation results to the ones produced by expert observers. There is also a lack of standardized statistical protocols for summarizing the results and making conclusions about algorithm performance. Finally, the evaluation study with expert-defined segmentations is tedious, time consuming and expensive to carry out.

Currently, in the field of image segmentation evaluation, a widely employed technique is to compare the automated segmentations from these algorithms to manual segmentation generated by a single expert (Gordon *et al* 2009, Udupa *et al* 2006, Petitjean and Dacher 2011). Metrics based on spatial distance are used when the delineation of the boundary is critical in the segmentation (Fenster and Chiu 2005). Also, in many evaluation studies (Skalski and Turcza 2011, Hadjiprocopis *et al* 2005, Wu and Jie 2003, Krishnamurthy *et al* 2004), the manual and automated segmentations are compared using a measure of region overlap (Crum *et al* 2006), such as the Dice coefficient (Dice 1945) and the Jaccard index (Jaccard 1912). Segmentation results are also evaluated by using manual segmentations obtained from multiple experts, using algorithms such as the simultaneous truth and performance level estimation (STAPLE) (Warfield *et al* 2004, Gordon *et al* 2009, Warfield *et al* 2008) algorithm.

In DWMRI, to evaluate the segmentation algorithms, the most widely employed technique is to compare the automated segmentation with a manual segmentation using a region-overlap approach (Krishnamurthy *et al* 2004, Wu and Jie 2003, Hadjiprocopis *et al* 2005, Jha *et al* 2010d). However, this approach suffers from various issues, apart from the general issues with image segmentation evaluation mentioned earlier. The manual segmentations are potentially error prone due to the low signal-to-noise ratio (SNR) of DW images, ghosting artifacts and fuzzy lesion boundaries (Krishnamurthy *et al* 2004, Jha 2009, Jha *et al* 2010d). Also, the process of acquiring segmentation results from experts is time consuming, expensive and tedious in DWMRI as well. Often, we have poor or, even worse, no manual segmentation results at all. In fact, in many cases, the radiologists do not perform manual segmentation on the DW images, but rather on a separate set of T1-weighted images that are acquired along with them. Therefore, the inaccuracy of the manual segmentation, or its absence, is a major issue in the evaluation of segmentation algorithms in DWMRI.

More significantly, in DWMRI, the images are acquired for a specific task, which is to compute the ADC value of the lesion. Lesion segmentation or any other intermediate image analysis algorithm is merely a step toward the end task of determining this ADC value. Therefore, an objective approach to evaluate the segmentation algorithms should also decide which algorithm aids the best in this task. The region-overlap methods do not evaluate the segmentation results based on this criterion. These methods are more suited for the task of determining whether the set of pixels that represent the object in manual segmentation also describe that object in the automated segmentation. While this criterion is important, and the segmentation algorithms should be evaluated based on this criterion, it is also important that they be evaluated based on the criterion of ADC estimation since that is the end task from these images. It was of interest for us to study if we could evaluate the segmentation algorithms in DWMRI on this task-based measure. This paper is an outcome of that study. The motivation for the study came from the work done in Barrett (1990), Barrett *et al* (1995, 1998), where the authors emphasize that an objective approach to image quality assessment must determine quantitatively how well the task required of the image can be performed from it. The purpose of this paper is to suggest a framework for task-based evaluation of segmentation algorithms in DWMRI in the absence of gold-standard segmentation.

Task-based evaluation of imaging systems or algorithms in the absence of a gold standard has been a technically challenging but important research problem. A major breakthrough for evaluating systems performing classification tasks in the absence of a gold standard was achieved by Henkelman *et al* (1990), who were able to perform receiver operating characteristic (ROC) analysis without knowing the true diagnosis. They demonstrated that the ROC parameters can be estimated by using two or more diagnostic tests, neither of which is accepted as the gold standard. A method to compare imaging systems for estimation tasks in the absence of a gold standard was developed by Hoppin *et al* (2002) and Kupinski *et al* (2002). The method was experimentally validated (Hoppin *et al* 2003) and used to compare ejection–fraction–estimation algorithms in the absence of a gold standard (Kupinski *et al* 2006). In this paper, we build upon the basic framework proposed in Hoppin *et al* (2002) and Kupinski *et al* (2002), to design a no-gold-standard technique to evaluate segmentation algorithms on the task-based measure of ADC estimation.

The basic idea of task-based evaluation of segmentation algorithms in DWMRI was proposed by us in Jha *et al* (2010c). In this paper, we carry out a significantly more rigorous study of the idea and the methods for its implementation. In the process, we suggest extensions to the original no-gold-standard approach so that it can be used more generally. The proposed technique compares the segmentation algorithms on the basis of both the ensemble mean

square error (EMSE) and precision. We also suggest a consistency check to verify whether the results obtained using the no-gold-standard approach are consistent with measured data.

2. Theory

2.1. ADC computation

Let us denote the b values at which the scan is performed by b_i, where i denotes the b value index. Assume we have P lesions each of which are imaged at two b values, b_1 and b_2, to give us P sets of images. In the pth set of images, the lesion is manually segmented to define the region of lesion pixels. Using the manual segmentation results for the pth set of images, the mean signal intensity of the lesion is calculated at both b values. Denote the mean signal intensities of the pth lesion at b values b_1 and b_2 by s^{m}_{p1} and s^{m}_{p2}, respectively, where the superscript m denotes manual segmentation. Also, denote the ADC of the pth lesion computed using the manual segmentation by a^{m}_p. The equation to compute a^{m}_p is given by (Theilmann *et al* 2004)

$$a^{\mathrm{m}}_p = \frac{-1}{b_1 - b_2} \ln\left(\frac{s^{\mathrm{m}}_{p1}}{s^{\mathrm{m}}_{p2}}\right). \tag{1}$$

Let the true ADC value of the pth lesion be a_p. Let there be K automated segmentation algorithms that we have to compare on the task-based measure of ADC estimation. Using the kth segmentation algorithm, we segment the lesion in the pth set of images. From the segmentation result, we obtain the mean signal intensity of the lesion at b values b_1 and b_2. Let us denote these mean signal intensities at the two b values by s^k_{p1} and s^k_{p2}, respectively. Using these mean signal intensities, we calculate the ADC of the lesion for the pth set, which we denote by a^k_p, as

$$a^k_p = \frac{-1}{b_1 - b_2} \ln\left(\frac{s^k_{p1}}{s^k_{p2}}\right). \tag{2}$$

We refer to the ADC value estimated using manual and automated segmentations as 'manual' and 'automated' ADCs, respectively, and denote them by the random variables A_m and A_k, respectively. The true ADC value is denoted by the random variable A.

2.2. Use of manual segmentations for task-based evaluation

The end task from DW images is ADC estimation, so the metric that ranks the segmentation algorithms on this task should quantify the error between the true and automated ADC values. To quantify this error for an estimation task, an appropriate performance metric is the EMSE (Barrett *et al* 2004, Whitaker *et al* 2008, Jha *et al* 2010c). The EMSE quantifies the error between the automated and true ADCs, averaged over the whole dataset of lesions, i.e. over different possible lesion variations and noise realizations. Therefore, the EMSE is a comprehensive figure of merit. For the kth automated segmentation algorithm, the EMSE, denoted by EMSE_k, is given by

$$\mathrm{EMSE}_k = E\{(A_k - A)^2\}, \tag{3}$$

where $E\{\ \}$ denotes the expected value of the quantity inside the parentheses. A segmentation algorithm is considered better if it has a lower value of EMSE_k. The issue is that in the absence of a gold-standard segmentation, we do not know the true ADC values, so that EMSE_k cannot be computed. However, we do know the manual ADC values. Our objective is to examine whether

the EMSE between the manual and automated ADC values, which we denote by $d(A_k, A_m)$, can serve as an indicator of the performance of the different segmentation algorithms. The expression $d(A_k, A_m)$ is mathematically defined as

$$d(A_k, A_m) = E\{(A_k - A_m)^2\}. \quad (4)$$

The expression can be rewritten by adding and subtracting the true ADC A inside the modulus sign, which yields

$$d(A_k, A_m) = E\{(A_k - A)^2\} + E\{(A - A_m)^2\} + 2E\{(A_k - A)(A - A_m)\}. \quad (5)$$

We denote $E\{(A - A_m)^2\}$, which is the EMSE between the true and manual ADC values, by $EMSE_{man}$. Using equation (3), the expression for $d(A_k, A_m)$ can be rewritten as

$$d(A_k, A_m) = EMSE_k + EMSE_{man} + 2E\{A_k(A - A_m)\} - 2E\{A(A - A_m)\}. \quad (6)$$

The rankings obtained using $d(A_k, A_m)$ and $EMSE_k$ will be the same if the expression for $d(A_k, A_m)$ is just the sum of $EMSE_k$ and terms that are independent of the kth segmentation algorithm. From the above expression, it is evident that this will occur only when the term $E\{A_k(A-A_m)\}$ vanishes. However, if there exists a systematic bias between the true and manual ADC values, then it can be easily shown that this term will not become zero. The manual segmentations themselves typically suffer from bias (Warfield *et al* 2008, 2004, Chalana and Kim 1997, Shirley *et al* 2011), which can easily lead to a bias between the true and manual ADC values. To study the existence of bias, we performed experiments with manual segmentations on simulated lesions with known ADC values. We detail on these experiments in section 4. The ADC values estimated from these manual segmentations provide further evidence of this bias. In light of these experimental observations and due to the documented presence of systematic bias between true and manual lesion segmentations, the assumption of absence of bias between true and manual ADCs is very debatable. Therefore, we cannot assume that the term $E\{A_k(A-A_m)\}$ will vanish, and thus, we cannot be sure if the ranking determined using $d(A_k, A_m)$ and $EMSE_k$ will be the same. As a result, using manual segmentations for evaluation of segmentation algorithms on this task-based measure and $d(A_k, A_m)$ as the figure of merit is not reliable.

2.3. The no-gold-standard approach

In this section, we propose the no-gold-standard approach to compare the automated segmentation algorithms in the absence of any manual segmentation results. As with manual segmentation, there could be bias between the true segmentation and the segmentation performed with any automated segmentation algorithm. Due to this bias, the automated ADC value will deviate from the true ADC value. We hypothesize that the relation between the true and automated ADC values should consist of a deterministic and a stochastic part. We model the deterministic part as a polynomial relationship between the true and automated ADC values. The motivation behind using a polynomial expression is that it can model any kind of relationship between the true and automated ADC values. The stochastic part is assumed to be a zero-mean normally distributed noise term. The reason for considering a zero-mean noise term is that if the noise is not zero-mean, then the deterministic part of the relationship will account for that using the bias term. We consider a normally distributed noise term since it is the most commonly used model for representing such errors. To further study the nature of this relationship, we performed a set of experiments, which we detail in section 4. From these experiments, we infer that a bias and a zero-mean normally distributed noise term are sufficient to relate the true and automated ADC values. Therefore, for the pth lesion, the

relation between the true ADC value and the ADC value estimated using the kth segmentation algorithm can be described through a bias term v_k and a noise term $\epsilon_p^k \sim \mathcal{N}(0, \sigma_k^2)$, as below:

$$a_p^k = a_p + v_k + \epsilon_p^k. \tag{7}$$

We assume that the parameters v_k and σ_k are characteristic of the kth segmentation algorithm and independent of the lesion. We would like to mention that although we are assuming the above linear relationship between the true and automated ADC values, our approach can be easily generalized to any general-polynomial relationship between these parameters.

The true ADC values are from different lesions and thus we assume that they have been sampled from a certain distribution. We choose this distribution to lie in the family of beta distributions since it is known that a beta distribution is flexible in modeling probabilistic data (Zou *et al* 2004) and can adapt itself to different probability distributions that the true ADC distribution can take. The beta distribution has the ability to model non-symmetric data, negatively skewed data as well as uni-modal, strictly increasing, strictly decreasing, concave, convex and uniform distributions, and therefore, is used in a wide range of applications (Romero 2010). Moreover, the beta distribution is the conjugate prior to a binomial distribution and has a potential for Bayesian extensions (Gelman *et al* 1995). The standard beta distribution returns values only between 0 and 1, but the true ADC value, a_p, has a different range of values, say (l, u), so that $0 \leqslant l \leqslant a_p \leqslant u$. However, this can easily be accommodated for by sampling the true ADC values from a four-parameter generalized beta distribution (GBD) (Johnson *et al* 1994, Wang 2011, Romero 2010, Karian and Dudewicz 2009) instead of a standard beta distribution. The four-parameter GBD, in addition to the two shape parameters ($\alpha > 0$ and $\beta > 0$), has parameters for the lower and upper limits of the distribution. It can be obtained by transforming the standard beta distribution using a recentering-rescaling transform (Johnson *et al* 1994). The probability density function of the four-parameter GBD of the true ADC values is

$$\mathrm{pr}(a_p|\alpha, \beta, g, l) = \frac{(a_p - l)^{(\alpha-1)}(g - a_p)^{(\beta-1)}}{B(\alpha, \beta)(g - l)^{(\alpha+\beta-1)}}, \tag{8}$$

where pr( ) denotes the probability density function of the quantity inside the parenthesis. Note that we do not know the values of $\{\alpha, \beta, g, l\}$, since we do not know the true ADC values or their prior distribution.

Our objective is to estimate the linear model parameters v_k and σ_k in equation (7) for each of the K segmentation algorithms, given only the automated ADC values a_p^k for the P lesions, and with no knowledge of the corresponding true ADC values a_p. This problem is thus equivalent to fitting the regression line without the x-axis. To solve this problem, we take a maximum-likelihood (ML) approach (Barrett *et al* 2004). Let the set of K noise terms, $\{\epsilon_p^1, \epsilon_p^2, ..., \epsilon_p^K\}$, for the K segmentation algorithms and the pth lesion be denoted by $\{\epsilon_p^k\}$. We assume that these noise terms are statistically independent, since they are due to different segmentations. Since we have modeled these noise terms to be normally distributed with zero-mean and standard deviation σ_k, with the above assumption of independence, we can write the joint probability density function of the noise terms $\{\epsilon_p^k\}$ for the pth lesion as

$$\mathrm{pr}\left(\left\{\epsilon_p^k\right\}\right) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_k{}^2}} \exp\left(\frac{-\epsilon_p^{k\,2}}{2\sigma_k{}^2}\right). \tag{9}$$

Using equation (7), we can rewrite the above equation as

$$\mathrm{pr}\left(\left\{a_p^k\right\}|\{v_k, \sigma_k\}, a_p\right) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_k{}^2}} \exp\left[\frac{-\left(a_p^k - a_p - v_k\right)^2}{2\sigma_k{}^2}\right], \tag{10}$$

where $\{a_p^k\}$ denotes the set of K ADC values estimated for the pth lesion using the K segmentation algorithms and $\{v_k, \sigma_k\}$ denotes the set of linear model parameters in equation (7) for all the K segmentation algorithms. Using the Bayes theorem and marginalizing equation (10) over a_p using equation (8), we obtain

$$\mathrm{pr}\left(\{a_p^k\}|\{v_k, \sigma_k\}, \{\alpha, \beta, g, l\}\right) = \int \mathrm{d}a_p \, \mathrm{pr}(a_p|\alpha, \beta, g, l) \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi {\sigma_k}^2}} \exp\left[\frac{-\left(a_p^k - a_p - v_k\right)^2}{2{\sigma_k}^2}\right]. \tag{11}$$

Since the different true ADC values are from different lesions, therefore we assume that the true ADC value a_p is statistically independent from one lesion to another. Using this assumption, the probability of all the observed automated ADC values, which we denote by the likelihood function L, can be written as

$$L = \prod_{p=1}^{P} \mathrm{pr}\left(\{a_p^k\}|\{v_k, \sigma_k\}, \{\alpha, \beta, g, l\}\right). \tag{12}$$

Taking the logarithm of both sides and using equation (11), we obtain the log-likelihood function

$$\lambda = \log(L) = P \log\left(\prod_{k=1}^{K} \frac{1}{\sqrt{2\pi {\sigma_k}^2}}\right) + \sum_{p=1}^{P} \log \int \mathrm{d}a_p \, \mathrm{pr}(a_p|\alpha, \beta, g, l) \exp\left[-\sum_{k=1}^{K} \frac{-\left(a_p^k - a_p - v_k\right)^2}{2{\sigma_k}^2}\right]. \tag{13}$$

Based on the philosophy of ML approach, we estimate the values $\{\{v_k, \sigma_k\}, \{\alpha, \beta, g, l\}\}$ for $k = 1, \ldots, K$ that maximize the probability of the observed data, or alternatively, the log-likelihood function λ given by equation (13). We denote the ML estimates of these parameters as $\{\{\hat{v}_k, \hat{\sigma}_k\}, \{\hat{\alpha}, \hat{\beta}, \hat{g}, \hat{l}\}\}$.

We use a quasi-Newton optimization technique (Coleman and Li 1994) in Matlab software to determine the values $\{\{\hat{v}_k, \hat{\sigma}_k\}, \{\hat{\alpha}, \hat{\beta}, \hat{g}, \hat{l}\}\}$ for which the maximum of the likelihood is obtained. We constrain this optimization to search between reasonable values of the parameters. To determine the search space for the beta distribution parameters, we assume that the automated and true ADC values would have similar distributions. Setting $v_k = 0$ and assuming no noise term in equation (7), we can obtain an approximate estimate of the true ADC distribution from the distribution of automated ADC values. To use the distribution of automated ADCs from the different segmentation algorithms, we first scale the histogram of automated ADC values so that all the automated ADC values lie between 0 and 1. For the kth segmentation algorithm, we determine the ML estimates of the beta distribution parameters $\hat{\alpha}_k$ and $\hat{\beta}_k$ that would fit these histograms (Hahn and Shapiro 1994). The search space for the beta distribution parameters α and β is restricted to lie between the minimum and maximum of the computed individual $\hat{\alpha}_k$ and $\hat{\beta}_k$ values. The search space for l and g is considered to be [0.5, 1.5] and [2.0, 3.0], respectively, in units of $\mathrm{mm}^2/10^3\mathrm{s}$ to encompass the typical range of ADC values encountered in our experiments. The search space for v_k and σ_k is kept as $[-0.5, 0.5]$ and [0.1, 1.0], respectively, again in units of $\mathrm{mm}^2/10^3\mathrm{s}$, to span a reasonably wide range of the bias and the standard deviation due to the noise term. The optimization routine is tuned to minimize the possibility of getting caught in a local minimum. To account for the possibility that the actual $\hat{\alpha}$ and $\hat{\beta}$ values might lie outside the specified search space for the beta parameters, we iterate the optimization routine. In each iteration of the routine, the search space for α and β parameters is increased. The iteration is repeated until the estimated

parameters or the likelihood function converge. Using the optimization routine, we estimate the values of $\hat{v}_k$ and $\hat{\sigma}_k$, i.e. the bias and variance in the automated ADC values, for each segmentation algorithm.

The bias and variance parameters estimated using the no-gold-standard approach can be used to rank the different segmentation algorithms on the basis of both the EMSE and precision. While the true EMSE cannot be computed since the true ADC values are not known. However, with our assumption regarding the model relating the true ADC values to the automated ADC values (equation (7)), we can derive an expression for the estimate of the EMSE between the true and automated ADC values. We denote this estimated EMSE for the kth segmentation algorithm by $\widehat{\text{EMSE}}_k$. Using equations (3) and (7), we can derive the expression for $\widehat{\text{EMSE}}_k$ to be

$$\widehat{\text{EMSE}}_k = \hat{v}_k^2 + \hat{\sigma}_k^2, \tag{14}$$

where we use the fact that ϵ_k is a zero-mean noise term. Using the estimated values of v_k and σ_k, $\widehat{\text{EMSE}}_k$ can be computed for the kth segmentation algorithm. A better performing algorithm will have a lower value of $\widehat{\text{EMSE}}_k$, and thus the algorithms can be ranked using this parameter.

To rank the algorithms on the basis of precision, we realize that the variance in the error of the automated ADC values is given by σ_k^2. This variance is inversely proportional to the precision of the ADC values obtained using a particular segmentation algorithm (Kupinski *et al* 2006). Therefore, the noise variance, or alternatively, the noise standard deviation value can be used to compare the segmentation algorithms on the basis of precision. The segmentation algorithm for which the value of the estimated noise standard deviation $\hat{\sigma}_k$ is the minimum is ranked as the most precise segmentation algorithm with respect to the task of ADC estimation using the no-gold-standard approach.

2.4. Consistency check

Using the suggested no-gold-standard approach, given the automated ADC values, we can rank the segmentation algorithms, but we cannot validate the computed rankings since we do not know the true ADC values. However, we can check whether the parameters obtained using the no-gold-standard approach are consistent with the estimated ADC values. Such consistency checks have been used to verify no-gold-standard approaches (Kupinski *et al* 2006). In this section, we suggest consistency checks for our proposed evaluation technique.

The consistency checks we suggest use the fact that if the parameters obtained using the no-gold-standard approach are consistent with the estimated ADCs, then the histogram of the estimated ADCs a_p^k should match the theoretically estimated distribution of a_p^k obtained using equation (7) for all the k segmentation algorithms. To plot the theoretically estimated distribution for a_p^k, we first generate a four-parameter GBD using the estimated beta distribution parameters $\{\hat{\alpha}, \hat{\beta}, \hat{g}, \hat{l}\}$ and the estimated bias term $\hat{v}_k$. We then convolve it with the normal distribution for the noise term with standard deviation $\hat{\sigma}_k$. The histogram of the automated ADC values a_p^k is then compared with this estimated distribution to check whether the two are similar. In order to evaluate the similarity of the two distributions, a quantile–quantile (Q-Q) plot (Wilk and Gnanadesikan 1968) is displayed. The Q-Q plot is a probability plot used for comparing probability distributions by plotting their quantiles against one another. If most of the values on this plot lie along the 45° line, then this indicates that the two distributions are similar, or equivalently, that the no-gold-standard output is consistent with measured data. To quantitatively evaluate how well the points lie along the 45° line, we compute the correlation coefficient between the quantiles of the two probability distributions. A correlation coefficient

close to 1 indicates that the two distributions are similar. We also perform Pearson's chi-squared test (Frieden 1991) to compare the similarity of the theoretically estimated distribution and the distribution of estimated ADCs. The X^2 test statistic is determined, and the p-value of this test statistic is computed. Based on the computed p-value, we can infer whether the two distributions are similar. We should mention that if these consistency checks pass, then that does not guarantee that the no-gold-standard technique has worked correctly, but its failure clearly indicates otherwise.

3. Materials and methods

3.1. In vivo imaging

In the study being carried out at the Arizona Cancer Center, DWMRI is being used to monitor the therapeutic response in breast cancer patients with metastases to the liver (Stephen *et al* 2011). Conventional T1- and T2-weighted imaging is performed at 1.5 T, along with DW single-shot echo-planar imaging (DW-SSEPI) using magnetic diffusion gradient values (b values) of 0 and 450 s mm^{-2}. Image parameters for the DW-SSEPI images are as follows: TE = 91.6 ms, 128 × 128 image matrix, FOV = 38 cm, TR = 6 s, BW = 250 kHz and 6 mm slice thickness. DWMRI image pairs at $b = 0$ and 450 s mm^{-2} are collected within a 24 s single breath hold. Each patient is imaged at day 0, 4, 11 and 39 following the commencement of cytotoxic therapy. From the *in vivo* study, we obtain a set of DW images for each lesion imaged. The lesions in the acquired images at different b values are segmented using a manual or automated segmentation algorithm. From the segmented lesion data, the mean signal intensity of the lesion at the different b values is computed, and using these data, the ADC value of the lesion is determined.

To verify our evaluation technique, it is ideal to use the real *in vivo* DW images of the liver containing the lesion. However, to determine the true ranking of the segmentation algorithms, we need to know the true ADC of the lesions, which we do not know in the real DW images. The next ideal approach is to develop bio-phantoms containing lesion-like structures having known ADC values and image them in a scanner. Although there has been research on designing phantoms with different ADC values (Laubach *et al* 1998), for this study we would require bio-phantoms with lesion-like structures that could be segmented with the different segmentation algorithms. Some recent works in this direction (Matsuya *et al* 2009, Matsumoto *et al* 2009), where the authors develop a bio-phantom containing tumor cells, are encouraging. However, the ADC values of these tumor cells are not known *a priori*, so again we do not know the true ADC value. Thus, we instead take the next-best strategy to verify our evaluation technique, which is using real DW images that contain simulated lesions with known ADC values. We now describe our approach to generate the simulated lesions for real DW images.

3.2. Lesion simulation

From our dataset of real patient images, we select seven sets of real images as templates for the background, where a set comprises of corresponding image slices at b values of 0 and 450 s mm^{-2}. Our next objective is to simulate lesions with known ADC value and insert them into the real DW images.

To simulate the lesion, we study the lesion images in the dataset. We select seven lesion images, all imaged at b value 0 s mm^{-2}, from our dataset. Each lesion is sufficiently different from the other in terms of size, intensity and other parameters. We observe that the shape of the lesions is approximately elliptical. The parameters of the ellipse are computed by determining

two values: the area of the lesion and the width–height ratio of the lesion. We also observe that the lesion intensity is almost constant, with some variability that we attribute to the presence of noise. This noise in MR images is Rician. However, in high-SNR regions, this noise can be approximated as Gaussian (Gudbjartsson and Patz 1995). We thus simulate the lesion as a constant intensity ellipse corrupted by zero-mean Gaussian noise. The constant intensity parameter and the standard deviation of the Gaussian noise for each lesion are assigned values equal to the mean signal intensity and the standard deviation of intensities of the corresponding real lesion. Using these parameters, we are able to simulate lesions at b value 0 s mm^{-2}.

For each lesion simulated at b value 0, we randomly select an ADC value from a normal distribution with mean 1.5 mm^2/(10^{3}s) and standard deviation of 0.3 mm^2/(10^{3}s). We choose this distribution to be a normal, and not beta distribution, since we want to validate if our method works even when the distribution of true ADC values is not a beta distribution. The mean and standard deviation of this normal distribution have been chosen based on the typical ADC values that we have observed in our study. Using this ADC value, the simulated lesion at b value 0 s mm^{-2} and the standard ADC equation, we simulate the lesion at b value 450 s mm^{-2}. For each lesion, we repeat this process ten times, each time having a different true ADC value and a different noise realization in the image. The lesions simulated at b values 0 and 450 s mm^{-2} are inserted into the real DW images at b values 0 and 450 s mm^{-2}, respectively. We compared the simulated images with the real DW images, and they were found to be very similar. At the end of this exercise, we have a dataset of 70 simulated DW images containing lesions with known ADC values. We generate two such datasets, which we denote by dataset 1 and dataset 2, respectively.

To study the performance of the method when the true ADC values are sampled from a beta distribution, we repeat the above procedure with the difference being that the true ADC value is sampled from a four-parameter GBD with $\alpha = 5$, $\beta = 5$, $l = 1.0\,\text{mm}^2/10^3\text{s}$ and $g = 2.0\,\text{mm}^2/10^3\text{s}$. These values for the parameters of this four-parameter GBD have again been chosen based on the typical ADC values that we observed in our study. Using the above procedure, we generate two more simulated-lesion datasets, that we denote as dataset 3 and dataset 4, respectively. Finally, we also generate another dataset of simulated lesions, in which the true ADC value is sampled from a normal distribution with mean 1.5 mm^2/(10^{3}s) and a standard deviation of 0.6 mm^2/(10^{3}s). The simulated-lesion dataset generated through this process is denoted as dataset 5.

3.3. Method to validate the no-gold-standard approach

The simulated lesion in the DW images is segmented manually and using three automated segmentation methods: a maximum-likelihood estimation (MLE) algorithm for segmenting lesions in digital mammograms (Kupinski and Giger 1998), a clustering algorithm (Pappas 1992) and an expectation–maximization (EM) algorithm (Zhang *et al* 2001). The segmentation algorithms are modified to perform segmentation only on a bounding box marked around the lesion as opposed to the entire image. From the segmentation results, we obtain the mean signal intensity of each lesion and thus compute the ADC of the lesion using equation (2) for each segmentation algorithm. This process is repeated for all the lesions in the two datasets.

The automated ADC values using the three segmentation algorithms for all the lesions are input to the no-gold-standard technique. The technique outputs $\{\hat{v}_k, \hat{\sigma}_k\}$ for the K segmentation algorithms. We first determine the rankings of the algorithm using the actual EMSE value. We can compute the actual EMSE since we *do* know the true ADC values in our simulation. To obtain the ranking using the no-gold-standard approach, as we mentioned earlier, we use $\widehat{\text{EMSE}}$ given by equation (14). If the rankings obtained using the two parameters are the

same, and the parameters are reasonably close to each other, it confirms that for our dataset, the no-gold-standard method is able to correctly rank the algorithms based on the EMSE parameter.

To determine the true ranking of the algorithms on the basis of precision, we use the residual sum of squares (RSS) as the figure of merit (Frieden 1991). The RSS is the measure of variance of the measured ADC values since the no-gold-standard approach is based on a linear-regression model. We evaluate the RSS by performing a least-squares (LS) fit of the true and automated ADC values, with the model as mentioned in equation (7). From the LS fit, the value for the bias parameter for the kth segmentation algorithm, $\hat{v}_{k,\mathrm{ls}}$ is obtained. The RSS for the kth segmentation algorithm is then given by

$$\mathrm{RSS}_k = \sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(a_p - a_p^k - \hat{v}_{k,\mathrm{ls}}\right)^2}. \tag{15}$$

The ranking for the no-gold-standard method is obtained using the noise variance $\hat{\sigma}_k$ estimated for each segmentation algorithm. We compare the two rankings and if they are similar, and to a certain degree, if the parameters RSS_k and σ_k are close, that helps to show that the no-gold-standard approach has successfully ranked the algorithms on the basis of precision for our simulated dataset.

4. Results

4.1. Validating the relationship between the true and estimated ADC values

On the simulated-lesion datasets 1 and 2, we use manual and different automated segmentation algorithms to compute the manual and automated ADC values. Our first set of experiments studies the relationship between the true ADC value and the measured ADC value estimated using the different segmentation techniques, including the manual segmentation technique.

Our first objective is to check for the existence of bias between the true and manual ADCs, A and A_{m}, respectively. To verify this, we perform the Student t-test (Frieden 1991) on the error between true and manual ADC values, i.e. $A - A_{\mathrm{m}}$, assuming that the error follows a normal distribution. The null hypothesis is that the mean of this distribution is zero, i.e. there is no bias. On performing the test, we find that this hypothesis can be rejected, as the p-value is less than 0.0001 for both the simulated-lesion datasets. Due to the presence of bias in manual ADC values, as we have demonstrated earlier, the parameter $d(A_k, A_{\mathrm{m}})$ cannot be used to rank the segmentation algorithms. We also perform the Student t-test for the difference between the true and automated ADC values, and the p-values obtained point to the existence of a bias between these ADC values also.

The next objective is to determine the polynomial order of the relationship between the true and measured ADC values. To determine this relationship, we first perform a LS fit between the true and estimated ADC values from each segmentation technique, for different polynomial orders. The different polynomial orders are the zeroth-order polynomial, a first-order polynomial with just the bias term, a first-order polynomial with bias and slope terms and second- and third-order polynomials. The LS curve fit is performed since in our model, the noise terms are normally distributed, and therefore, the polynomial coefficients determined by the LS method are the ML estimates of the coefficients (Barrett *et al* 2004). Next, to determine the polynomial relationship that best models the relation between true and automated ADC values, we use Akaike's information criterion (AIC) (Burnham and Anderson 2004). The AIC measures the relative goodness of fit of a statistical model by measuring the amount of information loss that occurs when a model is used to describe some measured data. Since we

Table 1. The Δ_i values for different polynomial orders, with different segmentation algorithms and with the simulated-lesion datasets 1 and 2.

		Polynomial order				
	Segm. Alg.	Zeroth order	First order, only bias	First order	Second order	Third order
Dataset 1	Manual	13.31	0	1.82	3.69	5.02
	Clustering	7.92	1.08	0	2.16	2.31
	MLE	0	2.10	4.30	6.56	8.52
	EM	13.61	0	2.05	4.24	6.23
Dataset 2	Manual	16.89	0	2.14	3.64	4.98
	Clustering	5.88	0	1.67	2.97	4.75
	MLE	0.81	0.42	0	2.07	3.37
	EM	1.79	0	2.08	3.90	5.61

have a finite amount of data, we use the AIC with the second-order correction term, denoted by $\mathrm{AIC_c}$. The $\mathrm{AIC_c}$ for LS estimation with normally distributed error is given by

$$\mathrm{AIC_c} = P\log(\delta^2) + 2Q + \frac{2Q(Q+1)}{P-Q-1}, \tag{16}$$

where Q is the number of parameters to be estimated, P is the number of lesions in the dataset and δ^2 is the RSS in the LS-fitted model. Using equation (16), we determine the value of AIC_i with different order polynomials, where the subscript i denotes the index for the polynomial. We then determine the minimum AIC_i, which we denote by $\mathrm{AIC_{min}}$. Subsequently, we rescale AIC_i by computing the term Δ_i given by (Burnham and Anderson 2004)

$$\Delta_i = \mathrm{AIC}_i - \mathrm{AIC_{min}}. \tag{17}$$

This transformation causes the best model to have $\Delta_i = 0$, while the rest of the models have positive values. We determine the set $\{\Delta_i\}$ for all the segmentation algorithms and also for the manual ADC values. This process is repeated for the second lesion dataset too. In table 1, the Δ_i values for the different segmentation algorithms with different polynomial orders for both the lesion datasets are presented. We observe that in most scenarios, the Δ_i value is equal to zero when the polynomial is first order, with only the bias term. Even when not zero, Δ_i is quite small when the polynomial is first order with only bias term, with its highest value being 2.1. As mentioned in Burnham and Anderson (2004), a model that has $\Delta_i \leqslant 2$ has substantial support in modeling the data. Based on the values obtained for Δ_i for the simulated dataset and the different segmentation algorithms, the first-order polynomial with no bias term is the most suitable for modeling the relation between the true and automated ADC values. This experiment helps to show that the model described using equation (7) is appropriate for the relation between true and automated ADC values.

4.2. Validating the no-gold-standard approach and consistency check

For all the simulated lesions in each simulated-lesion dataset, the automated ADC values using the three segmentation algorithms are computed, in units of $\mathrm{mm^2/10^3 s}$. These automated ADC values are input to the no-gold-standard method, which then estimates the bias and noise-variance parameters for each of the three segmentation algorithms. Using the bias and noise-variance parameters, the algorithms are ranked on the basis of the EMSE parameter and precision. The rankings using the no-gold-standard approach are then compared with the true rankings. Following this, the suggested consistency checks are performed, and it is

Table 2. True and estimated EMSE for the three segmentation algorithms.

	Dataset 1		Dataset 2	
Segm. alg.	True EMSE	Estimated EMSE	True EMSE	Estimated EMSE
Clustering	0.117	0.075	0.140	0.255
EM	0.469	0.361	0.769	0.870
MLE	0.646	0.744	0.474	0.458

Table 3. RSS and noise standard deviation for the three segmentation algorithms.

	Dataset 1		Dataset 2	
Segm. alg.	RSS	Noise std. dev.	RSS	Noise std. dev.
Clustering	0.315	0.246	0.357	0.289
EM	0.625	0.599	0.858	0.787
MLE	0.809	0.826	0.662	0.661

determined whether the consistency check output correctly predicts the success or failure of the no-gold-standard approach. The experiment is performed for all five simulated-lesion datasets. We now present the results for the individual datasets.

4.2.1. Datasets 1 and 2. Using the output obtained from the no-gold-standard method for datasets 1 and 2, the three automated segmentation algorithms are ranked on the basis of EMSE, as shown in table 2. As is evident from the observations, the rankings obtained using the true and estimated EMSE parameters are the same. The rank of the algorithms on the basis of precision is summarized in table 3. The no-gold-standard rankings obtained using the noise variance are the same as the true rankings obtained using the RSS parameter for both the simulated-lesion datasets.

On the basis of both EMSE and precision, the two simulated-lesion datasets yield different rankings for the three algorithms. This indicates a difference in performance of the algorithms on the different datasets. However, in spite of this, the no-gold-standard method predicts the same rankings as the true rankings for the two datasets, thus showing its efficacy.

Figure 1 shows the regression lines obtained using the estimated bias and variance parameters using the no-gold-standard approach, for the three segmentation algorithms for datasets 1 and 2, respectively. The true versus measured ADC values are also overlaid as a scatter plot for visual comparison. We note that although we have plotted the true ADC value on the x-axis, this information was not used to compute the regression-line parameters, which were computed entirely using the automated ADC values.

We then perform the suggested consistency checks. The Q-Q plot is plotted and is shown in figure 2. The correlation between the quantiles of the two distributions is also computed and is close to one in all the cases. Pearson's chi-squared test is performed to compare the theoretical and estimated distributions and in all the cases, the determined X^2 test statistic has a p-value close to 1, indicating that the distributions are very similar. Thus, the consistency-check output indicates that the no-gold-standard method has not failed, which is in accordance with the observation that the no-gold-standard method predicts the correct ranking of the segmentation algorithms.

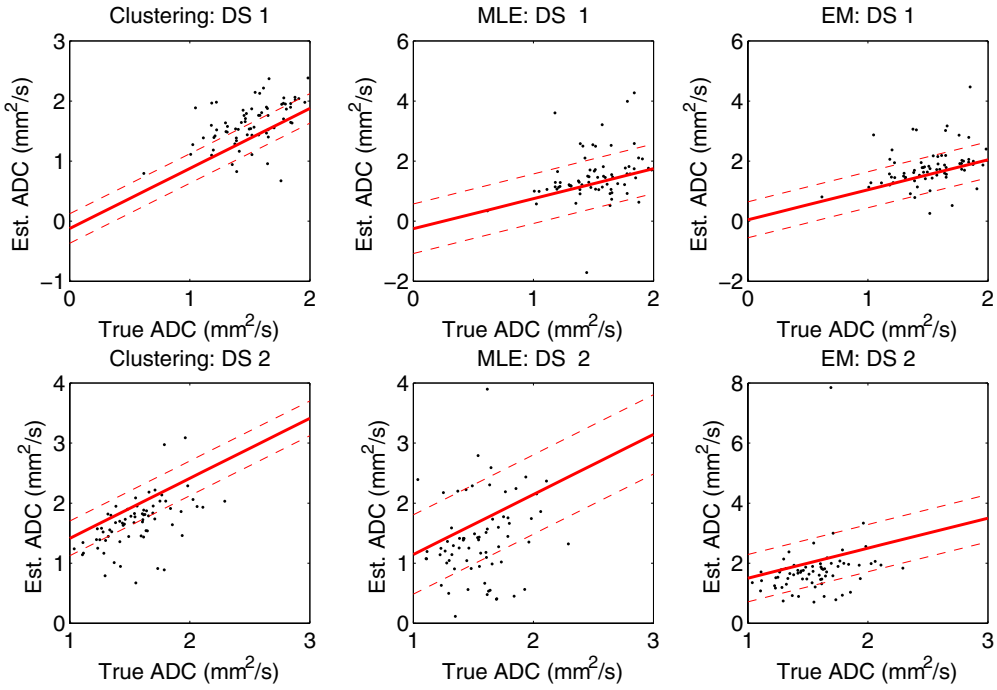

Figure 1. The regression lines estimated using the no-gold-standard method for the three segmentation algorithms and for the two simulated-lesion datasets: DS1 and DS2. In these two datasets, the true ADC values were sampled from a normal distribution. The solid line is generated using the estimated linear model parameters, and the dashed line denotes the estimated standard deviation. The scatter plots show the computed ADC values versus the true ADC value. ADC values are in units of $mm^2/10^3s$. Note that although we have plotted the true ADC value on the *x*-axis of the graph, this information was not used in computing the linear model parameters.

Table 4. True and estimated EMSE for datasets 3 and 4.

	Dataset 3		Dataset 4	
Segm. alg.	True EMSE	Estimated EMSE	True EMSE	Estimated EMSE
Clustering	0.056	0.106	0.101	0.211
EM	0.175	0.173	0.317	0.435
MLE	0.299	0.457	0.466	0.486

4.2.2. Datasets 3 and 4. The experiments are repeated for datasets 3 and 4. In these two datasets, as mentioned earlier, the true ADC values have been sampled from a four-parameter GBD. The algorithms are ranked on the basis of EMSE and precision, and as the results in tables 4 and 5 show, the no-gold-standard approach ranks the algorithms correctly. Figure 3 shows the regression lines obtained using the estimated bias and variance parameters for the three segmentation algorithms for these two datasets.

We next perform the suggested consistency checks. The Q-Q plot shown in figure 4, the correlations between the quantiles of the two distributions, which are close to 1 in all the cases, and Pearson's chi-squared test all indicate that the no-gold-standard approach has not failed.

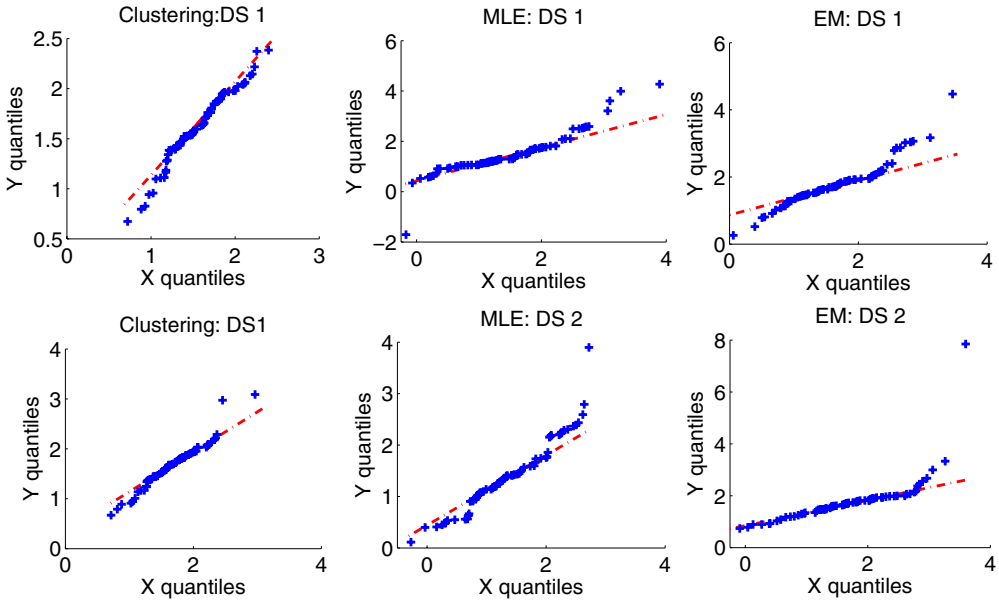

Figure 2. Consistency check: the Q-Q plot comparing the histogram of automated ADC values to the theoretically estimated distribution for the clustering, MLE- and EM-based segmentation algorithms and for the two simulated-lesion datasets: DS1 and DS2. The quantiles of the theoretically estimated distribution and the measured ADC distribution are plotted along the x- and y-axes, respectively. The 45° line is also plotted for visual convenience. We observe that the Q-Q plot lies approximately along the 45° line in all the scenarios, thus confirming that the measured and estimated distributions are consistent with each other.

Table 5. RSS and noise standard deviation for datasets 3 and 4.

	Dataset 3		Dataset 4	
Segm. alg.	RSS	Noise std. dev.	RSS	Noise std. dev.
Clustering	0.238	0.221	0.315	0.146
EM	0.408	0.371	0.547	0.430
MLE	0.506	0.505	0.645	0.675

This is again in accordance with the observation that the no-gold-standard method predicts the ranking of the segmentation algorithms correctly.

4.2.3. Dataset 5. On performing the experiment with dataset 5, the no-gold-standard approach does not rank the algorithms correctly. This is evident from the results shown in table 6, where we observe that the rankings from the no-gold-standard approach do not match the true rankings on the basis of both EMSE and precision. On performing the consistency checks, while the output using the Q-Q plot appears to be consistent with estimated ADC values, Pearson's chi-square test outputs a X^2 test statistic with a p-value close to 0 for two of the automated segmentation algorithms. These p-values are shown in the last column of table 6, and indicate that the no-gold-standard output is inconsistent with estimated ADC

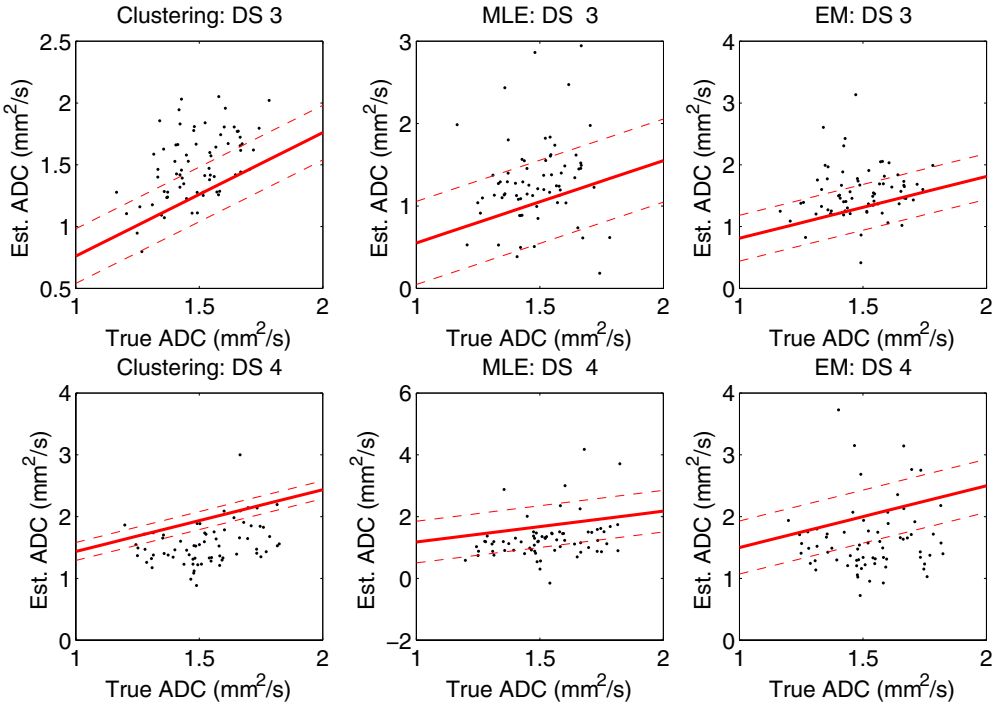

Figure 3. The regression lines estimated using the no-gold-standard method for the three segmentation algorithms and for the two simulated-lesion datasets: DS3 and DS4. The solid line is generated using the estimated linear model parameters, and the dashed line denotes the estimated standard deviation. The scatter plots show the computed ADC values versus the true ADC value. ADC values are in units of $mm^2/10^3 s$.

Table 6. No-gold-standard and consistency check result for dataset 5.

Segm. alg.	True EMSE	Estimated EMSE	RSS	Noise std. dev.	p-value of X^2 statistic
Clustering	0.419	0.169	0.546	0.295	1
EM	0.702	0.395	0.796	0.573	0
MLE	0.500	0.648	0.578	0.720	0

values. Thus, Pearson's chi-square consistency check is able to correctly predict the failure of the no-gold-standard approach for this dataset.

5. Discussion

In this paper, we have explored the idea of quantifying image analysis algorithms, especially image segmentation algorithms, based on how well they aid in achieving the end task from the image. In many image analysis tasks, the analyzed images are acquired for a certain purpose. This is especially true in medical imaging. Therefore, evaluating the image analysis algorithms on a task-based measure is important. Another idea explored in this paper is evaluating the segmentation algorithms in the absence of gold-standard or manual segmentation results. In most general images, and especially in medical images, achieving perfect manual segmentation

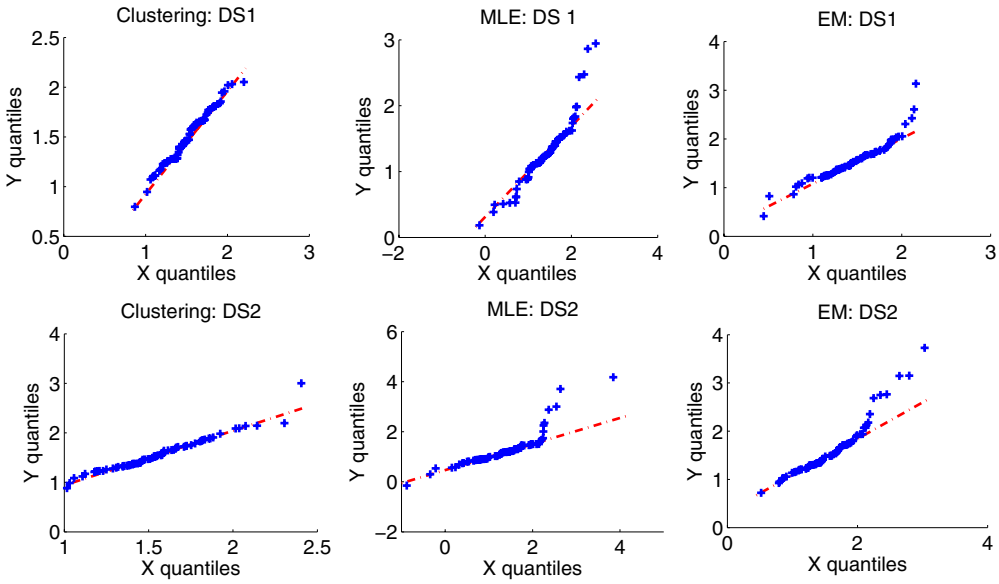

Figure 4. Consistency check: the Q-Q plot comparing the histogram of automated ADC values to the theoretically estimated distribution for the clustering, MLE- and EM-based segmentation algorithms and for the two simulated-lesion datasets: DS3 and DS4.

is almost impossible due to reasons discussed earlier. This paper recognizes this issue and proposes an evaluation method for DW images that accounts for this.

The desired task from DW images is accurate lesion-ADC estimation. We have shown using theoretical and experimental study that using manual instead of gold-standard segmentations to evaluate the automated segmentation algorithms is not reliable. We have then suggested a generalized-no-gold-standard approach to compare the automated segmentation algorithms. The no-gold-standard approach does not require either manual or perfect segmentation results and can rank the algorithms on the basis of both EMSE and precision. Apart from this, the estimated value of the bias can be used to rank the segmentation algorithms on the basis of accuracy. We have also suggested consistency checks to validate the output obtained from the no-gold-standard technique.

From experiments on our simulated-lesion dataset, we observed that even the manual ADC values are related to the true ADC values through the model described by equation (7). Therefore, the proposed no-gold-standard approach can also be used to rank multiple human experts performing manual segmentation. The bias and variance of the experts can be estimated. We can also compare the manual segmentation obtained by an expert to the outputs obtained using a set of automated segmentation algorithms, and therefore, decide which one is the best. Another advantage of our no-gold-standard approach is that it estimates the bias in the ADC values estimated using a segmentation algorithm. Therefore, using this bias value, the estimated ADC values can be corrected. Similarly, instead of segmentation algorithms, if human experts are performing the manual segmentation task, there may be a systematic bias in the segmentation performed by them (Warfield *et al* 2008). The no-gold-standard approach can be used to estimate and compensate for this bias. Apart from this, the no-gold-standard method also estimates the true distribution of the parameter of interest, i.e. the ADC value. This can provide information about the true distribution of the ADC values for the population being studied.

The original no-gold-standard approach (Hoppin *et al* 2002, Kupinski *et al* 2002) requires the parameter of interest to lie between 0 and 1. However, using the four-parameter GBD for the parameter of interest, we have extended the no-gold-standard approach to account for any range of the parameter of interest. Therefore, the scheme can be used in many other scenarios. For example, the segmentation algorithms to delineate tissue in single photon emission computed tomography images to determine the radioactive activity within the tissue (Floreby *et al* 1998) can be evaluated using the proposed method. Similarly, algorithms to segment the lesion in MR images of multiple sclerosis to compute its signal intensity (Rouaïnia *et al* 2006) can also be compared using the suggested no-gold-standard approach. However, in these applications, the model that relates the parameter of interest to the estimated value of the parameter should be carefully determined.

To determine the relation between the true- and automated-ADC values, we use simulated-lesion datasets. It would be ideal if this study could be done on bio-phantoms that have lesions with known ADC values. Based on some recent studies (Matsuya *et al* 2009), we predict that such a study could be possible in the near future. Also, it may be the case that, with a real-lesion dataset, the relation between true and estimated ADCs is more complicated than our model assumption. However, the suggested no-gold-standard method can be very easily adapted for that case, by just changing the polynomial order of the relationship. We would then have to estimate more coefficients of the model, than just the bias term. Also, the parameters based on which we rank the algorithms for accuracy and precision will change.

In this paper, we have focused on the performance of the suggested task-based technique, and not compared it with other conventional techniques like Dice's coefficient, Jaccard index and STAPLE (Warfield *et al* 2004). These evaluation techniques rank the segmentation algorithms based on region overlap. Thus, comparing our evaluation scheme, which ranks the algorithms based on how well they aid in the ADC estimation task, with the region-overlap-based schemes is inappropriate. In fact, the two approaches can yield different rankings, in which case, the diffusion analyst should decide the measure that is more critical for further study. In our opinion, measuring region overlap is also important to evaluate the efficacy of the segmentation algorithm, and thus, while evaluating segmentation algorithms, both the approaches should be used.

A requirement of the suggested no-gold-standard technique, and in fact of most other evaluation techniques, is the requirement of many lesion images. A large number of lesion images are required in the no-gold-standard technique since the number of parameters to be estimated is high. To compare three segmentation algorithms, we have to solve for ten parameters, i.e. $\{\{v_k, \sigma_k\}, \{\alpha, \beta, g, l\}\}$ for $k = 1, 2, 3$. It is appropriate if the dataset size is at least three times the number of parameters to be estimated. Generally, the datasets available are large, so this is not a very difficult requirement to satisfy.

The effect of outliers on the sensitivity of the no-gold-standard method is important to discuss. As we observe in the scatter plot shown in figures 1 and 3, some of the automated ADC values in our dataset are outliers. While few automated ADC values are negative, others are significantly greater than 3 $mm^2/10^3s$, which is the upper range of ADC values typically observed in our DWMRI study. The input to the no-gold-standard algorithm consists of all the automated ADC values, i.e. the outlier ADC values are not filtered from the dataset. Filtering the outlier automated values from the dataset will lead to unfair comparison between the segmentation algorithms, since if a segmentation algorithm outputs an outlier ADC value when the other algorithms do not, then such an output should negatively affect the rank of the particular segmentation algorithm. The no-gold-standard approach is able to rank the algorithms correctly in spite of the presence of these outlier ADC values for our dataset. The method is able to perform satisfactorily in the presence of some outlier values. However,

since it is a linear-regression technique, it will be sensitive to the presence of many outliers. Moreover, if a segmentation algorithm produces many outliers, then the assumption of a linear relationship between the true and automated ADC values can be violated. Therefore, if a segmentation algorithm consistently outputs poor segmentations or outlier ADC values, then it is not appropriate to use the no-gold-standard approach for its evaluation. In that case, it may perhaps be more useful to use a quantile–regression technique, which is less sensitive to the presence of outliers (Koenker and Hallock 2001).

Another important discussion topic is the validation of the output obtained using the no-gold-standard approach. The no-gold-standard method is essentially a ML estimation technique, and therefore, there can be error in the various parameters that are estimated using this technique. For example, we observe that the values of the estimated EMSE in the different experiments are not exactly the same as the true EMSE values. This error can affect the rankings that are determined using these estimated parameters, as we observe in dataset 5. In our experiments, we have not observed many cases where the error in parameter estimation has affected the rankings, but this possibility cannot be ignored. Therefore, the consistency checks that we have suggested are important and should be performed.

We would like to add that in our ADC-estimation procedure, we considered only a mono-exponential model for signal decay. However, in liver imaging, bi-exponential behavior has been observed over the range of b values used for clinical imaging (Taouli *et al* 2009). Also, the b values taken in this study can be subject to uncertainty due to tolerances in scanner gradient switching characteristics and timing limitations, but are assumed to be accurate, as is common in this type of analysis (Walker-Samuel *et al* 2009). A more robust approach to ADC estimation could take care of these issues as well. However, this will not have any effect on our no-gold-standard approach, as long as the true ADC values are related to the automated ADC values by just a bias and noise term. We do not expect that a small change in the ADC estimation technique would lead to any significant alteration of this relationship.

The rankings determined using the suggested no-gold-standard approach aid in the choice of the best segmentation algorithm for DW images of visceral organs. This segmentation algorithm can then be used in studies where lesion segmentation is required in the clinical DW images to determine the ADC value of the lesion. The computed ADC value can then be used to study the therapeutic response of a patient over a period of time, as is being done in various clinical studies (Theilmann *et al* 2004, Stephen *et al* 2011, Stephen *et al*). This can help in the growth of DWMRI as an accurate biomarker to predict anti-cancer therapy response, and lead to improved patient healthcare.

In this paper, we have dealt exclusively with the task of quantifying segmentation algorithms in DW images. However, there are other intermediate tasks that are performed to compute the ADC value from data acquired using DWMRI, such as image reconstruction (Sutton *et al* 2003) and ADC estimation (Jha *et al* 2010a, Jha and Rodriguez 2012). The algorithms for these intermediate steps can also be evaluated based on the no-gold-standard approach. More generally, there are many intermediate tasks that are performed to obtain the parameter of interest from an image, for which various algorithms can be developed. To compare these algorithms, task-based assessment can be used.

Acknowledgments

This work was supported in part by the National Institutes of Health, grant number NIH/NCI R01 CA119046, P30 CA23074 and RC1 EB010974. AKJ is partially funded by the technology research initiative fund (TRIF) imaging fellowship. The authors would like to thank Dr Harrison H Barrett for helpful discussions and Dr Eric Clarkson for reviewing initial drafts of

this paper. The authors would also like to thank Sundaresh Ram for help with the experiments and the anonymous reviewers for their valuable comments.

References

Bammer R 2003 Basic principles of diffusion-weighted imaging *Eur. J. Radiol.* **45** 169–84

Barrett H H 1990 Objective assessment of image quality: effects of quantum noise and object variability *J. Opt. Soc. Am.* A **7** 1266–78

Barrett H H, Abbey C K and Clarkson E 1998 Objective assessment of image quality: III. ROC metrics, ideal observers, and likelihood-generating functions *J. Opt. Soc. Am.* A **15** 1520–35

Barrett H H, Denny J L, Wagner R F and Myers K J 1995 Objective assessment of image quality: II. Fisher information, Fourier crosstalk, and figures of merit for task performance *J. Opt. Soc. Am.* A **12** 834–52

Barrett H H and Myers K J 2004 *Foundations of Image Science* 1st edn (New York: Wiley)

Bortner C D and Cidlowski J A 2002 Apoptotic volume decrease and the incredible shrinking cell *Cell Death Differ.* **9** 1307–10

Burnham K P and Anderson D R 2004 Multimodel inference *Sociol. Methods Res.* **33** 261–304

Chalana V and Kim Y 1997 A methodology for evaluation of boundary detection algorithms on medical images *IEEE Trans. Med. Imaging* **16** 642–52

Chenevert T L, Stegman L D, Taylor J M, Robertson P L, Greenberg H S, Rehemtulla A and Ross B D 2000 Diffusion magnetic resonance imaging: an early surrogate marker of therapeutic efficacy in brain tumors *J. Natl Cancer Inst.* **92** 2029–36

Coleman T F and Li Y 1994 On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds *Math. Prog.* **67** 189–224

Crum W R, Camara O and Hill D L 2006 Generalized overlap measures for evaluation and validation in medical image analysis *IEEE Trans. Med. Imaging* **25** 1451–61

Dice L R 1945 Measures of the amount of ecologic association between species *Ecology* **26** 297–302

Fenster A and Chiu B 2005 Evaluation of segmentation algorithms for medical imaging *Proc. IEEE Int. Conf. of the Engineering in Medicine and Biology Society* vol 7 pp 7186–9

Filippi M, Horsfield M A, Bressi S, Martinelli V, Baratti C, Reganati P, Campi A, Miller D H and Comi G 1995 Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis. A comparison of techniques *Brain* **118** 1593–600

Floreby L, Sjogreen K, Sornmo L and Ljungberg M 1998 Deformable Fourier surfaces for volume segmentation in SPECT *Proc. 14th Int. Conf. on Pattern Recognition* vol 1 pp 358–60

Frieden B R 1991 *Probability, Statistical Optics and Data Testing: A Problem Solving Approach (Springer Series in Information Sciences)* 3rd edn (Berlin: Springer)

Galons J P, Altbach M I, Paine-Murrieta G D, Taylor C W and Gillies R J 1999 Early increases in breast tumor xenograft water mobility in response to paclitaxel therapy detected by non-invasive diffusion magnetic resonance imaging *Neoplasia* **1** 113–7

Gelman A, Carlin J B, Stern H S and Rubin D B 1995 *Bayesian Data Analysis* (London: Chapman and Hall)

Gordon S, Lotenberg S, Long R, Antani S, Jeronimo J and Greenspan H 2009 Evaluation of uterine cervix segmentations using ground truth from multiple experts *Comput. Med. Imaging Graph.* **33** 205–16

Gudbjartsson H and Patz S 1995 The Rician distribution of noisy MRI data *Magn. Reson. Med.* **34** 910–4

Hadjiprocopis A, Rashid W and Tofts P S 2005 Unbiased segmentation of diffusion-weighted magnetic resonance images of the brain using iterative clustering *Magn. Reson. Imaging* **23** 877–85

Hahn G J and Shapiro S S 1994 *Statistical Models in Engineering* (New York: Wiley)

Henkelman R M, Kay I and Bronskill M J 1990 Receiver operator characteristic (ROC) analysis without truth *Med. Decis. Making* **10** 24–9

Hoppin J W, Kupinski M A, Kastis G A, Clarkson E and Barrett H H 2002 Objective comparison of quantitative imaging modalities without the use of a gold standard *IEEE Trans. Med. Imaging* **21** 441–9

Hoppin J W, Kupinski M A, Wilson D W, Peterson T E, Gershman B, Kastis G, Clarkson E, Furenlid L and Barrett H H 2003 Evaluating estimation techniques in medical imaging without a gold standard: experimental validation *Proc. SPIE* **5034** 230–7

Huisman T A 2003 Diffusion-weighted imaging: basic concepts and application in cerebral stroke and head trauma *Eur. Radiol.* **13** 2283–97

Jaccard P 1912 The distribution of the flora in the alpine zone *New Phytol.* **11** 37–50

Jha A K 2009 ADC Estimation in diffusion-weighted images *Master's Thesis* Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA

Jha A K, Kupinski M A, Rodriguez J J, Stephen R M and Stopeck A T 2010a ADC estimation of lesions in diffusion-weighted MR images: A maximum-likelihood approach *Proc. IEEE Southwest Symp. on Image Analysis and Interpretation* (*Austin, TX, USA*) pp 209–12

Jha A K, Kupinski M A, Rodriguez J J, Stephen R M and Stopeck A T 2010b ADC estimation in multi-scan DWMRI *Proc. Digital Image Processing and Analysis, OSA Technical Digest* (*Tucson, AZ, USA*) pp 1–3

Jha A K, Kupinski M A, Rodriguez J J, Stephen R M and Stopeck A T 2010c Evaluating segmentation algorithms for diffusion-weighted MR images: a task-based approach *Proc. SPIE* **7627** 76270L1–L8

Jha A K, Rodriguez J J, Stephen R M and Stopeck A T 2010d A clustering algorithm for liver lesion segmentation of diffusion-weighted MR images *Proc. IEEE Southwest Symp. on Image Analysis and Interpretation* (*Austin, TX, USA*) pp 93–6

Jha A K and Rodriguez J J 2012 A maximum-likelihood approach for ADC estimation of lesions in visciral organs *IEEE Southwest Symp. on Image Analysis and Interpretation* (*Tucson, AZ, USA*) pp 21–24

Johnson N L, Kotz S and Balakrishnan N 1994 *Continuous Univariate Distributions* vol 1 2nd edn (New York: Wiley)

Karian Z A and Dudewicz E J 2009 *Handbook of Fitting Statistical Distributions with R* (Boca Raton, FL: CRC Press)

Klauschen F, Goldman A, Barra V, Meyer-Lindenberg A and Lundervold A 2009 Evaluation of automated brain MR image segmentation and volumetry methods *Hum. Brain Mapp.* **30** 1310–27

Koenker R and Hallock K F 2001 Quantile regression *J. Econ. Perspect* **15** 143–56

Krishnamurthy C, Rodriguez J J and Gillies R 2004 Snake-based liver lesion segmentation *Proc. IEEE Southwest Symp. on Image Analysis and Interpretation* pp 187–91

Kupinski M A and Giger M 1998 Automated seeded lesion segmentation on digital mammograms *IEEE Trans. Med. Imaging* **17** 510–7

Kupinski M A, Hoppin J W, Clarkson E, Barrett H H and Kastis G A 2002 Estimation in medical imaging without a gold standard *Acad. Radiol.* **9** 290–7

Kupinski M A, Hoppin J W, Krasnow J, Dahlberg S, Leppo J A, King M A, Clarkson E and Barrett H H 2006 Comparing cardiac ejection fraction estimation algorithms without a gold standard *Acad. Radiol.* **13** 329–37

Laubach H J, Jakob P M, Loevblad K O, Baird A E, Bovo M P, Edelman R R and Warach S 1998 A phantom for diffusion-weighted imaging of acute stroke *J. Magn. Reson. Imaging* **8** 1349–54

Lu R, Marziliano P and Thng C H 2005 Liver tumor volume estimation by semi-automatic segmentation method *Proc. Int. Conf. IEEE Engineering in Medicine and Biology Society* pp 3296–9

Matsumoto Y *et al* 2009 *In vitro* experimental study of the relationship between the apparent diffusion coefficient and changes in cellularity and cell morphology *Oncol. Rep.* **22** 641–8

Matsuya R *et al* 2009 A new phantom using polyethylene glycol as an apparent diffusion coefficient standard for MR imaging *Int. J. Oncol.* **35** 893–900

Mohan V, Sundaramoorthi G and Tannenbaum A 2010 Tubular surface segmentation for extracting anatomical structures from medical imagery *IEEE Trans. Med. Imaging* **29** 1945–58

Muhi A, Ichikawa T, Motosugi U, Sano K, Matsuda M, Kitamura T, Nakazawa T and Araki T 2009 High-b-value diffusion-weighted MR imaging of hepatocellular lesions: estimation of grade of malignancy of hepatocellular carcinoma *J. Magn. Reson. Imaging* **30** 1005–11

Pappas T 1992 An adaptive clustering algorithm for image segmentation *IEEE Trans. Signal Process.* **40** 901–14

Petitjean C and Dacher J N 2011 A review of segmentation methods in short axis cardiac MR images *Med. Image Anal.* **15** 169–84

Romero A A 2010 Statistical adequacy and reliability of inference in regression-like models *PhD Thesis* Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Rouaïnia M, Medjram M S and Doghmane N 2006 Brain MRI segmentation and lesions detection by EM algorithm *Proc. World Academy Science Engineering Technology* vol 24 pp 139–42

Saad N M, Abu-Bakar S A R, Muda S and Mokji M 2010 Automated segmentation of brain lesion based on diffusion-weighted MRI using a split and merge approach *Proc. IEEE Conf. on Biomedical Engineering and Sciences* pp 475–80

Shirley J W, S Ty, Takebayashi S, Liu X and Gilbert D M 2011 FISH Finder: a high-throughput tool for analyzing FISH images *Bioinformatics* **27** 933–8

Skalski A and Turcza P 2011 Heart segmentation in echo images *Metrol. Meas. Syst.* **18** 305–14

Stephen R M *et al* 2011 Diffusion-weighted MRI of the liver: parameters of acquisition and analysis and predictors of chemotherapy response *Proc. Int. Society for Magnetic Resonance in Medicine (Montreal, Canada)* vol 19 p 1066

Stephen R M *et al* 2012 Diffusion-weighted MRI of the liver: parameters of acquisition and analysis and predictors of chemotherapy response *J. Mag. Res. Med.* (submitted)

Sutton B P, Noll D C and Fessler J A 2003 Fast, iterative image reconstruction for MRI in the presence of field inhomogeneities *IEEE Trans. Med. Imaging* **22** 178–88

Taouli B, Sandberg A, Stemmer A, Parikh T, Wong S, J Xu and Lee V S 2009 Diffusion-weighted imaging of the liver: comparison of navigator triggered and breathhold acquisitions *J. Magn. Reson. Imaging* **30** 561–8

Theilmann R J, Borders R, Trouard T P, Xia G, Outwater E, Ranger-Moore J, Gillies R J and Stopeck A 2004 Changes in water mobility measured by diffusion MRI predict response of metastatic breast cancer to chemotherapy *Neoplasia* **6** 831–7

Tunariu N, d'Arcy J A, Morgan V A, Germuska M, Simpkin C G, Giles S L, Collins D J and deSouza N M 2010 Assessment of variability of region of interest (ROI) delineation on diffusion weighted MRI (DW-MRI) using manual and semi-automated computer methods *ISMRM ESMRMB Joint Annual Meeting (Stockholm, Sweden)*

Udupa J K, Leblanc V R, Zhuge Y, Imielinska C, Schmidt H, Currie L M, Hirsch B E and Woodburn J 2006 A framework for evaluating image segmentation algorithms *Comput. Med. Imaging Graph.* **30** 75–87

Walker-Samuel S, Orton M, McPhail L D and Robinson S P 2009 Robust estimation of the apparent diffusion coefficient (ADC) in heterogeneous solid tumors *Magn. Reson. Med.* **62** 420–9

Wang J Z 2011 A note on estimation in the four-parameter beta distribution *Comm Stat.—Simul. Comput.* **34** 495–501

Warfield S K, Zou K H and Wells W M 2004 Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation *IEEE Trans. Med. Imaging* **23** 903–21

Warfield S K, Zou K H and Wells W M 2008 Validation of image segmentation by estimating rater bias and variance *Phil. Trans. A Math. Phys. Eng. Sci.* **366** 2361–75

Whitaker M K, Clarkson E and Barrett H H 2008 Estimating random signal parameters from noisy images with nuisance parameters: linear and scanning-linear methods *Opt. Express* **16** 8150–73

Wilk M B and Gnanadesikan R 1968 Probability plotting methods for the analysis of data *Biometrika* **55** 1–17

Wu L and Jie T 2003 Automatic segmentation of brain infarction in diffusion weighted MR images *Proc. SPIE* **5032** 1531–42

Zhang Y, Brady M and Smith S 2001 Segmentation of brain MR images through a hidden Markov random field model and the expectation–maximization algorithm *IEEE Trans. Med. Imaging* **20** 45–57

Zou K H, Wells W M, Kikinis R and Warfield S K 2004 Three validation metrics for automated probabilistic image segmentation of brain tumours *Stat. Med.* **23** 1259–82