Published in final edited form as:
Stat Med. 2004 April 30; 23(8): 1259–1282.

Three validation metrics for automated probabilistic image segmentation of brain tumours

Kelly H. Zou[1,2,*,†], **William M. Wells III**[1,3], **Ron Kikinis**[1], and **Simon K. Warfield**[1]

1 *Department of Radiology, Brigham and Women's Hospital, U.S.A.*

2 *Department of Health Care Policy, Harvard Medical School, U.S.A.*

3 *Artifcial Intelligence Laboratory, Massachusetts Institute of Technology, U.S.A.*

SUMMARY

The validity of brain tumour segmentation is an important issue in image processing because it has a direct impact on surgical planning. We examined the segmentation accuracy based on three two-sample validation metrics against the estimated composite latent gold standard, which was derived from several experts' manual segmentations by an EM algorithm. The distribution functions of the tumour and control pixel data were parametrically assumed to be a mixture of two beta distributions with different shape parameters. We estimated the corresponding receiver operating characteristic curve, Dice similarity coefficient, and mutual information, over all possible decision thresholds. Based on each validation metric, an optimal threshold was then computed via maximization. We illustrated these methods on MR imaging data from nine brain tumour cases of three different tumour types, each consisting of a large number of pixels. The automated segmentation yielded satisfactory accuracy with varied optimal thresholds. The performances of these validation metrics were also investigated via Monte Carlo simulation. Extensions of incorporating spatial correlation structures using a Markov random field model were considered.

Keywords

sensitivity; specificity; receiver operating characteristic (ROC) curve; dice similarity coefficient (DSC); mutual information; expectation maximization (EM) algorithm

1. INTRODUCTION

1.1. Image segmentation

Surgical planning and image-guided intervention procedures increasingly employ automated segmentation algorithms. MR imaging of the brain provides useful information about its anatomical structure, enabling and facilitating quantitative pathological or clinical investigation. Brain segmentation is a useful image processing method [1]. It assigns unique labels to two or more classes, e.g. skin, brain tissue, ventricles, and tumour, representing an anatomic structure to each pixel in an input grey-level image [2–4].

*Correspondence to: Kelly H. Zou, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A..
†E-mail:zou@bwh.harvard.edu

This work was supported by NIH grants R01LM7861, P41RR13218, P01CA67165, R01RR11747, R01CA86879, R21CA89449-01, R03HS13234, a research grant from the Whitaker Foundation, and a New Concept Award from the Center for Integration of Medicine and Innovative Technology.

Binary (i.e. two-class) manual segmentation is the simpliest, most frequently employed, and yet time-consuming procedure. It also has the additional difficulty of differentiating subtle intensity variations among pixels, particularly for those on the border of a tumour. However, the results of such manual segmentations, i.e. outlining a tumour by hand, is stilled often regarded as the gold standard. When performed by experts, such 'truth' may ultimately influence the amount and degree of tumour removal, and thus is a critical step in clinical practice.

Recently, Warfield *et al.* [5,6] have proposed an automated segmentation algorithm that yields continuous pixel-wise probabilistic measures indicative of pixel-wise malignancy, with an application to brain tumours [6]. It assigns 0 as non-malignancy to 1 as malignancy. Thus, the appropriate methodology for validating this and similar continuous segmentation algorithms are required, which motivated our investigation. Our goal was to examine several validation metrics to compared the fractional segmentation against multiple experts' manual segmentations as the gold standard.

1.2. Gold standard

The most important element in validating the accuracy of a segmentation algorithm is the gold standard. Particularly, in the image segmentation problem, it is the classification truth of each pixel. For simplicity, let us assume a two-class truth by labelling the non-tumour class as C_0 and tumour class as C_1. Medical images typically present a challenge with a large number of pixel data available, for example, $N = 256 \times 256 = 65\,536$ in a grey-scale image. The fraction of target malignant class varies depending on the type of the disease. In this application, the brain tumours occupied about 10–20 per cent of the entire brain, which still comprised a large number of pixels within the tumour regions [7]. However, one may also construct a 'region of interest' (ROI), a subset of the entire brain pixels, which contains the target tumour of interest. The size of the ROI can vary, depending on the specific image processing task and the clinical information included in this region.

1.3. Summary accuracy measures

It is common to assume that all pixels are independent. In discriminant imaging analysis, mixture models such as normal, normal mixture, or histograms are used. For example flexible histogram approach [8,9], classifications of several target features may be conducted by treating pixels as a set of independent samples drawn from a mixture distribution [10]. A distribution of feature vectors using Parzen windows was modelled. Despite a simplification by ignoring the natural pairing induced by pixels and the dependence between pixels in close proximity, the independence approaches, i.e. essentially analyses of histograms, are frequently employed in image processing and explorative analyses. These methods tend to give reasonable results and are much less computationally intensive than those assuming spatial correlation structions (see Section 1.5 for methods to incorporate spatial information).

For the purpose of comparing two sets of binary segmentation results, summary statistics for the analysis of contingency tables may be adopted [11]. A χ^2 test of independence may be conducted between segmentations and the corresponding binary gold standard values. To evaluate the agreement between the binary segmentation and the gold standard, one may also compute the kappa statistic [11,12]. By considering spatial information, Jaccard (JSC) [13] and Dice (DSC) [14] similarity coefficients are typically used as a measure of overlap, where DSC ranges from 0, indicating no spatial alignment between these two sets of binary segmentation results, to 1, indicating complete alignment. In addition, boundary measures such as the Hausdorff distance between segmentation and the gold standard, may be computed [15].

To evaluate the performance of a 'continuous classifier', the distributions in the two distinct classes, here labelled as C_0 and C_1, respectively, may directly be compared using the classical two-sample statistics with appropriate assumptions. A popular visual method for assessing the overall classification accuracy is to plot a receiver operating characteristic (ROC) curve, a function of sensitivity vs (1-specificity). See several review articles on the ROC methodology [16–18]. Previously, we have developed several ROC methods including non-parametric, semi-parametric and parametric transformation models, for estimating and comparing ROC curves derived from continuous data [19–21]. Regression methodology has also found in the literature [22–25]. To utilize spatial correlation in the image feature space, medical image classification applications have used techniques, e.g. the co-occurrence matrix (CM) developed by Haralick *et al*. [26], to enhance image features and to capture texture information. A CM is a probability density distribution of two pixel intensities, conditioned on distance and angle between the two pixels. Other methods for incorporating spatial information are given in Section 1.5.

Because a large number of pixels are typically present, a test of statistical significance may not be meaningful clinically. We suggest avoiding relying on statistical hypothesis testing, along with the use of *p*-values, which fails to convey important quantitative information. Thus, in this article, hypothesis testing over pixels is not conducted. Instead of focusing on the resulting statistical significance based on *p*-values, we confine attention on estimation of several summary accuracy measures as metrics of valid segmentation.

However, hypothesis testing may still be conducted in a subset of the entire image based in an ROI. Statistical hypothesis tests include a Student's *t*-test, and a non-parametric Mann–Whitney *U*-statistic [27–29]. Alternatively, a Kolmogorov–Smirnov test may be used to directly compare the two underlying distributions [30]. To consider spatial information between the two sets, we have recently developed hypothesis testing methods using the method of CM in prostate imaging segmentations [26,31]. We scaled 256 × 256 image intensity data to fit a range between 0 and 255 by mapping these 256 levels into the range of [max{0, (mean − 3 SD)},(mean + 3 SD)]. For each centre pixel, the CM matrix was constructed by a number of neighbours that were equidistant from the centre pixel. To model the texture differences of the tissues of classes, log likelihood ratios, $\log\{\Pr(C_1)/\Pr(C_0)\}$, with C_0 for non-prostate tumour pixels and C_1 for prostate tumour pixels, for each of these neighbouring pairs. We used the median of the log-likelihoods as the feature statistic; $\Pr(C_1)$ and $\Pr(C_0)$ were the conditional probabilities of tumour and non-tumour pixels, respectively, obtained from the CM entries that correspond to the pair of pixel intensities and distance between pixels. Because image interpreters often considered the slice above or below the current image slice for confirmation of prostate cancer, we have also extended the CM method to three-dimension by constructing CM one slice above and below the current slice.

1.4. Optimal threshold

The optimal threshold, or operating point, is of importance in developing guidelines for clinical decision-making. Optimal operating threshold value may be computed using decision approaches or via optimization. For example, one may simultaneous maximize a function of predicting a sensitivity and a specificity (see Section 4.4) over a range of thresholds [32–35]. Alternatively, one may consider maximizing a function of these accuracy values or their utilities, along with the prevalence of the disease and cost issues associated with misclassifications of both positive and negative outcomes [36,37].

1.5. Incorporation of spatial correlation

There exist several solutions, which incorporate spatial correlation structures: For example, an image segmenter may correctly identify a target object, but unfortunately may locate it

incorrectly. Thus, the task is not only a classification but also a localization problem. In the diagnostic literature, an extension called LROC was predicted from either a binary or multi-category ROC, as well as the number of possible locations of the LROC [38–40]. The *y*-axis is the proportion of responses both correctly detected and localized, while the *x*-axis is the false positive rate.

Second, a spatially varying prior probability of the target class, C_1, may be adopted to reflect *a priori* about the spatial information. For example, when assessing brain segmentations of white matter from other classes (e.g. grey matter, cerebral spinal fluid and fat), a probabilistic atlas [41,42] of the distribution of white matter, derived from a large group of subjects, can provide the spatial knowledge of the location at each pixel. Clustered samples may be taken from regions of the image, and statistical parametric inferences may be made. However, frequently, such an atlas is not available for many structures of interest.

Alternatively, a Markov random field (MRF) prior may be used, as in this work. The MRF model assumes that the spatial information in an image is encoded through contextual constraints of neighbouring pixels. The constraints make it more likely that the neighbouring pixels belong to the same class. Expectation-maximization (EM) algorithm is often used to derive the summary statistics of each labelling class [43–45]. The MRF method will be briefly considered.

1.6. Overview of the methodology

The goal of this work is to develop three accuracy metrics, ROC curve, mutual information, and Dice similarity coefficient for validating automated probabilistic brain tumour segmentations. Our methods are developed mainly under the pixel independence assumptions. Brief extensions to spatial correlation are also made.

This article is organized as follows. In Section 2, we introduce notations for the classification problem. The composite gold standard is estimated using an EM algorithm in Section 3. In the same section, mixture modelling will be developed for continuous probabilistic segmentation, with parameters estimated by matching-moments. In Section 4, we examine three different functions of these estimated parameters, which are subsequently used as the validation metrics for evaluating image classification accuracy. Based on each metric, an optimal threshold is recommended by maximizing the likelihood function. The performance of these validation metrics is investigated via Monte Carlo simulation in Section 5. Section 6 presents a clinical example of MRI of three different types of brain tumours, with these validation metrics illustrated on nine brain tumour cases. Finally, discussion and extensions to include pair-wise spatial correlation structures using an MRF model are provided in Section 7.

2. NOTATIONS AND ASSUMPTIONS

For simplicity, we assume that individual pixels belong to one of two distinct and independent populations (i.e. non-tumour control class, C_0 vs tumour class, C_1), determined by the gold standard (truth), T. Consider two random samples, X_i ($i = 1,...,m$) and Y_j ($j = 1,...,n$), drawn from C_0 and C_1, respectively. The observed continuous random variable is labelled Z, representing the probabilistic pixel-wise segmentation measures. Note that the domain of Z is [0,1]. This continuous random variable Z generates our probabilistic segmentation data, while the gold standard T determines the true pixel-wise classes. Stratified by the gold standard, for each member of class C_0, there is a measurement $X \sim (Z|T = 0)$ assumed to have cumulative distribution function (c.d.f.) F, with probability density function (p.d.f.) f and survival function $\bar{F}=1-F$. Similarly, for each member of class C_1, there is a measurement $Y \sim (Z|T = 1)$ assumed to have c.d.f. G with p.d.f. g and survival function $\bar{G}=1-G$.

We assume that the gold standard, T, has a Bernoulli distribution, with a probability of $\Pr(T = 0) = \pi = m/(m + n)$ for class C_0, and the tumour probability of $\Pr(T=1)=\bar{\pi}=1-\pi=n/(m+n)$ for class C_1. By Bayes' Theorem, the marginal distribution of Z is a mixture of F and G, with mixing proportions π and $\bar{\pi}$. That is, c.d.f of $H=\pi\cdot F+\bar{\pi}\cdot G$ with p.d.f. h, where the p.d.f. of Z is

$$h(z) = \pi f(z) + \bar{\pi} g(z) \quad \text{with} \quad \pi + \bar{\pi} = 1 \quad (\forall z \in [0,1]) \tag{1}$$

Specifying any arbitrary threshold, $\gamma \in (0,1)$ for Z yields a discretized version of a decision random variable, D_γ. This implies the equivalence of the following events: $\{D_\gamma = 0\} \equiv \{Z \leq \gamma\}$ and $\{D_\gamma = 1\} \equiv \{Z > \gamma\}$. Thus, a 2×2 contingency table (Table 1) is constructed.

3. A MIXTURE MODEL

3.1. Estimation of the gold standard

Instead of directly observing the gold standard, T, we conduct manual segmentations by a total of R expert readers, each performing binary manual segmentation labelled B_{lr} ($l = 1,...,N = m+n$; $r = 1,...,R$). Such repeated reading will enable us to estimate a combined gold standard, using the missing data approach similar to those found in the literature [46,47].

Let Q_{0r} and Q_{1r} represent the true accuracy rates under the classes, C_0 and C_1, respectively. The inter-expert decisions are assumed to be conditionally independent, given the latent truth. We only observe binary classification decision B_{lr}, i.e. $(B_{lr}|T_l, Q_{0r}, Q_{1r}) \perp (B_{lr'}|T_l, Q_{0r'}, Q_{1r'})$, for any two different experts, $r \neq r'$.

We now wish to estimate the latent vector $\mathbf{T}$, of length N, by $\hat{\mathbf{T}} = \arg\max_T \Pr(\mathbf{B}|\mathbf{T},\mathbf{Q_0},\mathbf{Q_1})$, for all $N = m + n$ pixels. However, these segmenter-specific classification qualities, $\mathbf{Q_0}$ and $\mathbf{Q_1}$ each a vector of length R are unknown quantities. To estimate the pixel-wise gold standard, we have developed a software program named 'Simultaneous Truth and Performance Level Estimation' (STAPLE) [48,49] using the following iterative ($k = 1,...,K$) EM algorithm [50, 51]:

The expectation (*E*) *step*: In the $(k-1)$th iteration ($k = 1,...,K$ till convergence is reached), let

$$u^{(k-1)} = \prod_{r:B_{lr}=1} \widehat{Q}_{1r}^{(k-1)} \prod_{r:B_{lr}=0} (1-\widehat{Q}_{1r}^{(k-1)}) \quad \text{and} \quad \upsilon^{(k-1)} = \prod_{r:B_{lr}=0} \widehat{Q}_{0r}^{(k-1)} \prod_{r:B_{lr}=1} (1-\widehat{Q}_{0r}^{(k-1)})$$

Define the weight variable in the common notations for EM algorithms [50,51] based on the $(k-1)$th iteration:

$$w_l^{(k-1)} = f(T_l = 1 \,|\, B_l, \widehat{Q}_{0r}^{(k-1)}, \widehat{Q}_{1r}^{(k-1)}) = u^{(k-1)}\bar{\pi}/(u^{(k-1)}\bar{\pi} + \upsilon^{(k-1)}\pi)$$

where $\pi = \Pr(T = 0)$, $\bar{\pi}=\Pr(T=1)$, and $\pi+\bar{\pi}=1$, as given before. At $k = 0$, the initial estimates of the Q_{0r}'s and Q_{1r}'s may be based on a voting scheme to drive the gold standard, using a majority rule over $\mathbf{B}$.

The maximization (M) step: At the kth iterative step, to maximize the log-likelihood,

$$(\widehat{Q}_0^{(k)}, \widehat{Q}_1^{(k)}) = \arg \max_{(Q_0,Q_1)} E[\log \{f(B\,|T, Q_0, Q_1) f(T)\} \,|\, f(T \,|\, B, \widehat{Q}_0^{(k-1)}, \widehat{Q}_1^{(k-1)})]$$

and that for the rth segmenter,

$$(\widehat{Q}_{0r}^{(k)}, \widehat{Q}_{1r}^{(k)}) = \arg \max_{(Q_{0r}, Q_{1r})} \sum_{l:B_{lr}=0} \left\{ \overline{w}_l^{(k-1)} \log Q_{0r} + w_l^{(k-1)} \log(1-Q_{1r}) \right\} + \sum_{l:B_{lr}=1} \left\{ \overline{w}_l^{(k-1)} \log(1-Q_{0r}) + w_l^{(k-1)} \log Q_{1r} \right\}$$

where wl(k-1) is the weight variable from the $(k-1)$th iteration, and w¯l(k-1)=1-wl(k-1).

The MLEs of the accuracy quality parameters are, respectively:

$$\widehat{Q}_{0r}^{(k)} = \frac{\sum_{l:B_{lr}=0} \overline{w}_l^{(k-1)}}{\sum_{l:B_{lr}=0} \overline{w}_l^{(k-1)} + \sum_{l:B_{lr}=1} \overline{w}_l^{(k-1)}} \quad \text{and} \quad \widehat{Q}_{1r}^{(k)} = \frac{\sum_{l:B_{lr}=1} w_l^{(k-1)}}{\sum_{l:B_{lr}=1} w_l^{(k-1)} + \sum_{l:B_{lr}=0} w_l^{(k-1)}}$$

In our experience, typically only $K < 20$ iterations were necessary till convergence. More details on this algorithm have been described in separate articles [48,49]. Relevant software codes are available from the authors.

3.2. Modelling of probabilistic segmentation data

Recall that the continuous random variables, X and Y, are the probabilistic segmentation results for classes C_0 and C_1, stratified by the gold standard T, respectively. Because both X and Y take values between [0,1], it is conventional and flexible to assume independent beta distributions, i.e. $F(x) \sim \text{Beta}(\alpha_0, \beta_0)$ and $G(y) \sim \text{Beta}(\alpha_1, \beta_1)$. It is known that a beta distribution is flexible in modeling probabilistic data and is conjugate to a binomial distribution, and thus has a potential for Bayesian extensions [52]. In the simulation study presented later, we will consider several mixtures of beta distributions.

Due to a large number of pixels in image processing, instead of estimating the parameters of the beta distributions by their iterative MLEs [53], we use a matching-moment approach. Usually there is a lack of efficiency when using the method of moments; however, the loss of efficiency is small or negligible with a large number of pixels available in our problem. Furthermore, the lack of efficiency for the method of moments does not relate to bias.

The expectation and variance of a Beta(α, β) distribution are given by α/(α+β) and αβ/{(α +β)^2 (α + ± 1)}, respectively. Thus, the estimates (α^0,β^0) of the shape parameters based on the **x**-sample of C_0 may be obtained by matching the first two moments (mean and variance).

In order to match the sample mean x¯ and standard deviation s_x of the **x**-sample, it can be shown that

$$\widehat{\alpha}_0 = \overline{x}\left\{\overline{x}(1-\overline{x})/s_x^2 - 1\right\}, \quad \text{and} \quad \widehat{\beta}_0 = (1-\overline{x})\left\{\overline{x}(1-\overline{x})/s_x^2 - 1\right\}$$

Similarly for α^1 and β^1, computed based on the two moments, y¯ and $\underline{s}_y$, of the **y**-sample of C_1. We now present three validation metrics in the following section, with a higher value in [0,1] indicating higher accuracy.

4. VALIDATION METRICS

4.1. Sensitivity, specificity, and ROC curve

The accuracy of a diagnostic test can be summarized in terms of an ROC curve [16–18]. It is a plot of sensitivity (true tumour fraction) vs (1-specificity) (true non-tumour fraction) based on Z and T, at all possible thresholds.

Conventionally, $p\gamma=\bar{F}(\gamma)$ is labelled as false positive rate (FPR or 1-specificity), on the *x*-axis of an ROC curve. True positive rate (TPR or sensitivity) is $q\gamma=\bar{G}(\gamma)$ at the specified γ, or $qp=\bar{G}\circ\bar{F}\text{-}1(p)$ at any specified *p*, on the *y*-axis of an ROC curve. The ROC curve is given by $(\bar{F}(\gamma),\bar{G}(\gamma))$ for $\gamma\in[0,1]$, or $(p,\bar{G}\circ\bar{F}\text{-}1(p))$ for $p\in[0,1]$. There is always a trade-off between these two false positive and false negative error rates, or specificity and sensitivity.

An overall summary accuracy measure is the area under the ROC curve (AUC):

$$\mathrm{AUC} = P(X < Y) = \int_{\gamma=0}^{1} \bar{G}(\gamma)\, \mathrm{d}\bar{F}(\gamma) = \int_{p=0}^{1} q(p)\, \mathrm{d}p \tag{2}$$

4.2. Dice similarity coefficient

At any arbitrary threshold γ, the Dice similarity coefficient [14], DSC_γ, may be computed as a function of the sensitivity and specificity. Following the convention of an ROC plot, here we label the false positive rate $p_\gamma = P(Z > \gamma|T = 0)$ and the true positive rate $q_\gamma = P(Z > \gamma|T = 1) = P(D_\gamma = 1|T = 1)$. According to the Bayes' Theorem, the Jaccard similarity coefficient at γ, JSC_γ, is first defined as the pixel ratio of union and intersection between the two tumour classes determined separately by D_γ and T [13]:

$$\begin{aligned}
\mathrm{JSC}_\gamma &= \frac{\#\{(D_\gamma=1)\cap(T=1)\}}{\#\{(D_\gamma=1)\cup(T=1)\}} \\
&= \frac{P(D_\gamma=1\,|\,T=1)P(T=1)}{P(D_\gamma=1)+P(T=1)-P(D_\gamma=1\,|\,T=1)P(T=1)} \\
&= \frac{P(D_\gamma=1\,|\,T=1)P(T=1)}{P(D_\gamma=1\,|\,T=0)P(T=0)+P(T=1)} \\
&= \frac{\bar{\pi}q_\gamma}{\pi p_\gamma+\bar{\pi}} \\
&= \frac{\bar{\pi}\bar{G}(\gamma)}{\pi\bar{F}(\gamma)+\bar{\pi}}
\end{aligned}$$

Note that as we tend to confine our attention to the spatial overlap in the tumour class, DSC for the non-tumour may be computed analogously, but is generally not be of interest. A generalized overall DSC over all threshold levels is defined as a simple function of JSC_γ and by integrating out γ [14], by assuming a uniform distribution for γ:

$$\mathrm{DSC} = \int_{\gamma=0}^{1} 2(\mathrm{JSC}_\gamma)/(\mathrm{JSC}_\gamma + 1)d\gamma \tag{3}$$

4.3. Entropy and mutual information

The mutual information (MI) between the binary decision D_γ at any threshold γ and the gold standard T can be computed as follows [54]:

$$\mathrm{MI}_\gamma = \mathrm{MI}(D_\gamma, T) = H(D_\gamma) + H(T) - H(D_\gamma, T) \tag{4}$$

where

$$\begin{aligned}
H(D_\gamma) &= -(p_{11} + p_{12})\log_2(p_{11} + p_{12}) - (p_{21} + p_{22})\log_2(p_{21} + p_{22}) \\
&= -(\pi\bar{p}_\gamma + \bar{\pi}\bar{q}_\gamma)\log_2(\pi\bar{p}_\gamma + \bar{\pi}\bar{q}_\gamma) - (\pi p_\gamma + \bar{\pi}q_\gamma)\log_2(\pi p_\gamma + \bar{\pi}q_\gamma)
\end{aligned}$$

$$\begin{aligned}
H(T) &= -(p_{11} + p_{21})\log_2(p_{11} + p_{21}) - (p_{12} + p_{22})\log_2(p_{12} + p_{22}) \\
&= -\pi\log_2(\pi) - \bar{\pi}\log_2(\bar{\pi})
\end{aligned}$$

$$\begin{aligned}
H(D_\gamma, T) &= -p_{11}\log_2(p_{11}) - p_{12}\log_2(p_{12}) - p_{21}\log_2(p_{21}) - p_{22}\log_2(p_{22}) \\
&= -\pi\bar{p}_\gamma\log_2(\pi\bar{p}_\gamma) - \bar{\pi}\bar{q}_\gamma\log_2(\bar{\pi}\bar{q}_\gamma) - \pi p_\gamma\log_2(\pi p_\gamma) - \bar{\pi}q_\gamma\log_2(\bar{\pi}q_\gamma)
\end{aligned}$$

with the joint probabilities, (p_{11}, p_{12}, p_{21}, p_{22}), given in Table I.

MI between the continuous random variable Z and T may also be computed using a conditional entropy approach:

$$\begin{aligned}\mathrm{MI}(Z,T) &= H(Z)-H(Z|T)\\ &= -E_Z[\log_2\{k(Z)\}]-\pi E_Z[\log_2\{f(Z)\}]-\bar{\pi}E_Z[\log_2\{g(Z)\}]\\ &= -\int_{z=0}^{1}[k(z)\log_2\{k(z)\}-\pi f(z)\log_2\{f(z)\}-\bar{\pi}g(z)\log_2\{g(z)\}]\,dz\end{aligned}$$

where $k(z)=\pi f(z)+\bar{\pi}g(z)$ as in (1).

4.4. Determination of an optimal threshold

The above accuracy criteria may be maximized numerically over the entire range of γ to derive an optimal threshold, γ^opt, as the recommended operating cut-off point [32]. Because there is always a trade-off between sensitivity and specificity (that is, when sensitivity is 1, specificity is 0, and vice versa).

Note that the expression for the area under the ROC curve in (2) is free of γ after being integrated out. Thus, we only illustrate the square root of the sum of squares of sensitivity and specificity values as our optimization criterion.

Other functions, such as the mean accuracy values of sensitivity and specificity may also be considered for maximization. Furthermore, γ may be computed by maximizing the mutual information MI_γ, and Dice similarity coefficient (DSC_γ).

5. A SIMULATION STUDY

5.1. Designs

A Monte Carlo empirical simulation study was first undertaken to examine the estimated validation metrics by simulating a region-of-interest with varied pixels. The following realistic values were fixed for this study in the context of image processing with typically a large number of pixels in a region of interest. Under each combination of sample size and distributional assumptions, 500 random samples were drawn repeatedly with data generated in the following two experiments:

Experiment 1: The effect of the size of a region of interest: The size of the brain tumour was first realistically fixed at n = 1000 for class C_1. However, the background non-tumour pixels in an ROI was varied at m = {1000; 3000; 9000}, for class C_0, implying that the target tumour of fixed size occupied anywhere in $\bar{\pi}$ = {50 per cent; 25 per cent; 10 per cent}. In effect, we increased the size of an ROI. This is clinically useful, as sometimes features and structures other than the tumour alone in the ROI may provide additional clinical information in performing segmentations.

Experiment 2: The effect of the size of a tumour: The total size of the brain, excluding extra surrounding background, was fixed at $N = m+n$= 10 000, for class C_0 and C_1 combined. However, the proportion of the tumour in the entire brain pixel varied. The numbers of pixels were assumed to be (m,n) = {(8500,500); (9000,1000); (8500,1500)}, thus, the proportions of tumour occupied $\bar{\pi}$ = {15 per cent; 10 per cent; 5 per cent} of the brain. In effect, the tumour size decreased within the brain of fixed size.

The effect of the beta mixture models employed in both experiments: Under each combination of m and n, the beta mixture distributions were generated using the following combinations of four shape parameters for classes C_0 and C_1 under the equal variances assumption,

$(\alpha_0,\beta_0,\alpha_1,\alpha_1) = \{(1, 1, 1, 1); (1, 1\ 5, 1\ 5, 1); (1, 2, 2, 1); (1, 2\ 5, 2\ 5, 1); (1, 3, 3, 1); (1, 9, 9, 1)\}$; as well as under the unequal variances assumption, {(1,3,1.5, 1); (1, 1.5, 3, 1); (1, 9, 3, 1); (1, 3, 9, 1)}.

See Figure 1 for these various beta distributions of class C_0, with the given shape parameters, where $\alpha_0 < \beta_0$. The distributions under class C_1 may be graphed similarly, with $\alpha_1 > \beta_1$. The distributions are symmetric about $x = 0.5$ if $\alpha_0 = \beta_1$ and $\beta_0 = \alpha_1$. For example, The p.d.f.'s of Beta(1,3) and Beta(3,1) are symmetric about $x = 0.5$, with means 0.25 and 0.75, respectively. The specified distributions corresponded to the scenarios that the diagnostic accuracy ranged from flipping a coin (i.e. uniform distributions when the beta shape parameters were all equal to 1, with AUC = 0.5) to very high (i.e. skewed distributions towards the extremes of 0 and 1, with AUC ≈ 1).

5.2. Results

Experiment 1: The effect of the size of a region of interest: Tables II and III show the estimated validation metrics under the combinations when tumour size was fixed at 1000 pixels the varied ROI size. We observed that the biases of the estimates of the beta parameters were small.

The estimated AUC ranged from 0.500 to 1.000; MI ranged from 0.000 to 0.992; DSC ranged from 0.152 to 0.889 over all of the combinations of sample size and distributional assumptions.

All accuracy metrics varied if the distributional assumptions differed. The mean values of these metrics increased if the separation between the two beta distributions under c_0 and C_0 became more apparent.

We observe that among these three metrics, the means of the AUC values were quite robust with respect to the choice of the non-uniform beta mixtures. In addition, they were fairly constant regardless of the size of the ROI.

MI, on the other hand, was sensitive to both the distributional assumption and particularly the size of the ROI. Even given under the same distributions, it varied dramatically over different ROI sizes. For example, MI was 0.465 when the tumour occupied only 10 per cent of the ROI of a total size of 10 000, but was improved to 0.992 when it occupied 50 per cent of the ROI of the same total size.

The DSC was influenced somewhat by the sample sizes, but more so by the distributional assumptions.

This suggested that perhaps segmentation tasks are better performed if an ROI surrounding the target tumour can be pre-determined as close as possible to the actual location of the tumour. However, sometimes it would be difficult to crop the image into a rather small ROI if the original brain image contains additional important and useful information and features, in which case it may not be reasonable to perform segmentation tasks within a restricted ROI.

Experiment 2: The effect of the size of a tumour: Tables IV and V show the estimated values of the validation metrics under the combinations when the total brain ROI size was fixed at 10 000 while the proportion of the tumour varied. We observed that the biases of the estimates of the beta parameters were also negligible.

The estimated AUC ranged from 0.500 to 1.000; MI ranged from 0.000 to 0.604; DSC ranged from 0.152 to 0.800, over all of the sample size and distributional assumptions considered.

We demonstrated that, again, the means of the AUC values were robust in terms of the choice of the non-uniform beta mixtures and the size of the ROI.

MI tended to take on low values (0.283–0.604) even under extremely skewed beta-distributions with parameters $(\alpha_0;,\beta_0,\alpha_1,\beta_1) = (1, 9, 9, 1)$, especially for tumours that occupied no more than 15 per cent of the brain.

The DSC was influenced somewhat by both sample size and by the distributional assumptions, and the values were intermediate, i.e. lower than the corresponding AUCs but higher than the MIs.

This suggested that the segmentation method works better for larger tumours in the brain. Therefore, for very smaller tumours, perhaps it is sensible to define an ROI in advance rather than using the entire brain image, as in Experiment 1, in order to increase the relative size of the tumour to the background prior to performing either manual or automated segmentation tasks.

6. A CLINICAL EXAMPLE

6.1. Materials and methods

The cases: A total of nine patients were randomly selected from a neuro-surgical database of 260 brain tumour patients, of which three had meningiomas (M), three astrocytomas (A), and three other low-grade gliomas (G) [7,55]. Visually the meningiomas enhanced better on grey-scale images than the remaining two tumour types.

Imaging protocol: Patient heads were imaged in the sagittal planes with a 1.5T MR imaging system (Signa, GE Medical Systems, Milwaukee, WI), with a postcontrast 3D sagittal spoiled gradient recalled (SPGR) acquisition with contiguous slices (flip angle, 45°); repetition time (TR), 35 ms; echo time (TE), 7 ms; field of view, 240 mm; slice-thickness, 1:5 mm; 256 × 256 × 124 matrix). The acquired MR images were transferred onto a UNIX network via Ethernet.

Automated probabilistic segmentation: The automated probabilistic segmentation was the relative tumour probability of lesion per pixel with signal intensity modelled as a Gaussian mixture of the two classes, based on an initial semi-automated binary segmentation [7].

Manual binary segmentation and gold standard estimation: An interactive segmentation tool (MRX, GE Medical Systems, Schenectady, NY) was employed and ran on an Ultra 10 Workstation (Sun Microsystems, Mountain View, CA). The structures were outlined slice-wise by expert operators using a mouse on a Sun Workstation. The program connected consecutive points with lines. An anatomical object was defined by closed contour, and the program labelled every pixel of the enclosed volume.

For the purpose of validation, we randomly selected one single 2D slice for each case from the subset of the MR volume with the tumour. Manual segmentation was performed independently by three experts (blinded to the machine segmentation results) to outline the brain and the tumour. The pixel-wise composite gold standard was determined by the EM Algorithm STAPLE [48,49]. The remaining pixels were labelled as background. Stratified analyses are conducted by case and tumour type.

Statistical computing: All computations and optimizations were performed on a SunMicrosystem SunBlade 100 Workstation and in Matlab6, S-Plus6.0 and C languages.

6.2. Results

As an example case, we show the grey scale MR image of a meningioma case 1 (Figure 2). The corresponding semi-automated binary segmentations of this image is also displayed

(Figure 3). It was then used to derive the probabilistic results. The resulting empirical and approximated beta mixture densities are plotted in Figure 4.

For all cases, we now report in Table VI the pixel counts (*m;n*), and the relative proportion of the tumour size, stratified by the ground truth. The sample means and SDs of the non-tumour and tumour probability data are reported. These sample moments were used to estimate the shape parameters of the beta mixture distributions. As shown in this table, the proportion and size of the tumours varied anywhere between 2.6 and 16.3 per cent. In general, the three low-grade gliomas tended to have larger number of pixels with higher proportions. The estimated beta distributions were variable and were skewed under both non-tumour and tumour classes, evidenced by the unequal α and β parameters as they would under symmetry.

The overall validation accuracies presented in Table VII were generally high but were variable. The estimated ROC curves for all cases by tumour type are displayed in Figure 5. Based on AUC and DSC, the probabilistic segmentation algorithm was most accurate for meningiomas but least for astrocytomas. The accuracy was more variable among the astrocytoma cases. Although MI appeared low in about half of these cases but were still above 0, suggesting a possible reason that these brain tumours were quite small compared to the entire brain.

In the same Table VII, the recommended optimal thresholds would depend heavily on the metric criterion used for optimization. One may suggest basing the function of sensitivity and specificity as the criterion if both quantities are important. If the spatial consideration of the 'tumour class' is more important, then the threshold based on the overlap statistics would be appropriate. However, an overall recommended metric is AUC, while the operating threshold would be based on the MI as an optimization criterion.

7. DISCUSSION

In this work, we have presented systematic approaches to validating the accuracy of automated image segmentation results leading to pixel-wise probabilistic interpretation of the tumour class. We developed an EM algorithm for estimating the latent gold standard. In addition, we modelled the probabilistic segmentation results using a mixture of two beta distributions with different shape parameters. Summary accuracy measures, including ROC curve, mutual information, and Dice similarity coefficient, were estimated. An optimal threshold was also derived under each metric.

The example data generally showed satisfactory accuracy based on our automated segmentation algorithm, particularly illustrated by the high AUC and DSC values. In addition, the estimated MI values were above zero even for small tumours relative to the entire brain. The recommended optimal threshold, however, seemed to be case- and task metric-dependent. Thus, for different optimization purposes, one may consider different clinical meaningful threshold as their operating point of choice, and the basis for such optimization depends largely on the clinical goal.

The main advantage of our approaches is that the parametric beta mixture modelling for continuous automated probabilistic segmentation data was simple and probabilistic. The estimation procedures for the beta shape parameters are straightforward, fast and unbiased, using the moment-based approach. In addition, we derived a composite gold standard by an EM algorithm, which was essential for many validation purposes. It is known that there frequently exists inter- and intra-observer variability even among or within experts' manual segmentation results. Thus, the proposed methods may be generalizable to similar statistical validation tasks of combining segmentation methods. In addition, these metrics may be used in the context of assessing general diagnostic tests or classification algorithms.

Our recommendation as to which metric to use is as follows: When the overall classification accuracy is of interest, an ROC analysis may be appropriate. When a more refined index examining the accuracy, which is sensitive to changes in the sizes of the tumour and number of pixels, the mutual information should be adopted. Finally, when spatial alignment and configuration are of interest, then the overlap measure should be considered. Depending on the different validation purposes, the appropriate optimal threshold can be recommended. Such selections of the appropriate metrics should ultimately reflect the relevant clinical task and disease type. Unfortunately, because of these varied tasks, the optimal thresholds tend to differ, based on the particular validation metrics employed as the optimization criteria.

We have assumed an independence model, and spatial information may be incorporated using a pair-wise MRF model (see the appendix). Clinically, most segmentation errors occur at the interface between two structures. The i.i.d. assumption in each class may also be a simplification although such an assumption is often used and is easy to interpret. Nevertheless, the lack of independence between scores in different pixels would result in different estimates. The MRF modelling in the appendix is rather computationally intensive. Unfortunately, in our example little difference was observed under the independence and MRF models in terms of the pixel-wise gold standard estimates. In the particular meningioma case illustrated in Figure 6, the non-MRF and MRF gave very similar results. If we were to binarize the estimated gold standard, then there would not be differences in the estimated boundaries of this tumour. This is partly due to having only three segmentations, and the very smooth boundary of the tumour.

To further investigate the complexity of the interior and surface of the tumour, we have conducted a preliminary analysis using logistic regression to examine the effects of the distance between the pixel and the centre of a tumour the angle, and the probabilistic segmentation result, to predict the gold standard. Encouragingly, our result showed that the automated segmentation method performed well even on the boundary of the tumour in increased distance away from the centre. Moreover, different intra-cluster correlations may be present, separately for pixels on the border and for pixels in the middle of a tumour, and thus should be incorporated in the validation.

The above issues are beyond the scope of this article but will further be invested in future research without assuming pixel-wise independence by adapting the Ising model, cluster analysis, and the MRF model [56,57]. We plan to develop an hierarchical approach to estimate the validation metrics by incorporating both pixel-level, cluster-level (e.g. border vs centre), and even higher-level (e.g. tumour type) covariate information.

Finally, apart from the Monte Carlo simulation experiments conducted in Section 5, a digital brain phantom with simulated image data placing realistic tumours and possibly additional complex anatomical structures will be constructed to evaluate the performances of our validation metrics.

In summary, the proposed three validation metrics can provide a simple way to evaluate and validate a continuous segmentation or classification algorithm in imaging processing and analysis. More sophisticated statistical methodology is called for to deal with complex imaging data as encountered frequently in medical diagnosis and treatment planning.

APPENDIX A: INCORPORATION OF A SPATIAL PRIOR

A pair-wise MRF model may be employed to incorporate spatial homogeneity. We have extended our EM algorithm in the STAPLE program for estimating the pixel-wise gold standard T_l as a more refined standard to validate the automated algorithm on.

At each pixel l, we assume that the segmenter quality estimates are independent of the underlying gold standard, (Q^0r,Q^1r)⊥Tl. Then an MRF prior is added by assuming pair-wise interactions between all neighbouring pixels l and l' ($l \neq l'$; $\{l; l'\} = 1, ...,N$) in order to induce spatial homogeneity.

Maximization (*M*) *step with an MRF prior*: Following closely the notations of Greig [56], at the kth iteration the log-likelihood of the pixel-wise gold standards becomes

$$\log f(T \mid B, \widehat{Q}_0^{(k-1)}, \widehat{Q}_1^{(k-1)}) \propto \sum_l \sum_{l'} \eta_{ll'} \{T_l T_{l'} + (1-T_l)(1-T_{l'})\} + \sum_l T_l \cdot \log \left\{ \frac{\Pi_r f(B_{lr} \mid T_l=1, \widehat{Q}_{0r}^{(k-1)}, \widehat{Q}_{1r}^{(k-1)}) f(T_l=1)}{\Pi_r f(B_{lr} \mid T_l=0, \widehat{Q}_{0r}^{(k-1)}, \widehat{Q}_{1r}^{(k-1)}) f(T_l=0)} \right\}$$

where $\eta_{ll'} \geq 0$ and encodes a spatial prior indicating an interaction weight between neighbouring pixels l and l'. If $\eta_{ll'} > 0$, then these neighbouring pixels have spatial homogeneity; $\eta_{ll'} = 0$ implies pixel-wise independence. In practice, $\eta_{ll'} = \eta$, for adjacent pixels ($l;l'$). The estimate of T_l is efficiently made using a (maximum flow)−(minimum cut) network flow problem [56].

In Figure 6, a comparison between the estimated gold standard with the frequencies of three manual segmenters' results is provided (left panel). The MRF modelling (right panel) is presented for the meningioma Case 1, using the spatial prior $\eta_{ll'} = \eta = 2.5$. The independence model described earlier and the MRF models yielded slight differences in probabilities of the presence of tumour under the MRF model to that under the independence model, particularly in regions on the boundary of the tumour. However, such differences were not perceptible in the index scale. In the future, we will further investigate the differences and incorporate such MRF modelling in the three validation metrics, which will be beyond the scope of this work.

Acknowledgements

We acknowledge with thanks image data from Dr Michael R. Kaus of the Philips Research Laboratories, as well as the constructive comments from Drs Ferenc Jolesz and Steven Haker of the Surgical Planning Laboratory of Brigham and Women's Hospital. We thank the three experts who performed manual segmentations of the brain tumour cases.

References

1. Pham DL, Xu CY, Prince JL. Current methods in medical image segmentation. Annual Review of Biomedical Engineering 2000;2:315–325.
2. Cline HE, Lorensen, Kikinis R, Jolesz F. Three-Dimensional segmentation of MR images of the head using probability and connectivity. Journal of Computer Assisted Tomography 1990;14:1037–1045. [PubMed: 2229557]
3. Rogowska J. Overview and fundamentals of medical image segmentation. *Handbook of Medical Imaging* Academic Press: San Diego, 2000, 69–85.
4. Goldszal AF, Pham DL. Volumetric segmentation. *Handbook of Medical Imaging* Academic Press: San Diego, 2000, 185–211.
5. Warfield SK, Kaus MR, Jolesz FA, Kikinis R. Adaptive template moderated spatially varying statistical classification. *Proceedings of the 1st International Conference on Medical Image Computing and Computer-Assisted Intervention* Boston, MA, 1998, 431–438.
6. Warfield SK, Westin CF, Guttmann CRG, Albert M, Jolesz FA, Kikinis R. Fractional segmentation of white matter. *Proceedings of 2nd International Conference on Medical Imaging Computing and Computer Assisted Interventions*, Springer: Cambridge, U.K., 1999.
7. Kaus M, Warfield SK, Nabavi A, Black PM, Jolesz FA, Kikinis, R. Automated segmentation of MRI of brain tumours. Radiology 2001;218:586–591. [PubMed: 11161183]
8. De Bonet JS, Viola P, Fisher JW III. Flexible histograms: a multiresolution target discrimination model. In Zelnio EG (ed.), *Proceedings of SPIE*, vol. 3370, 1998.

9. De Bonet JS, Viola P. Texture recognition using a non-parametric multi-scale statistical model. *Proceedings of Computer Vision and Pattern Recognitions Conference* 1998.
10. Santago P, Gage HD. Quantification of MR brain images by mixture density and partial volume modeling. IEEE Transaction on Medical Imaging 1993;12:566–574.
11. Fleiss JL. *Statistical Methods for Rates and Proportions.* Wiley: New York, 1981, 212–236.
12. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Transactions on Medical Imaging 1994;13:716–724. [PubMed: 18218550]
13. Jaccard P. The distribution of flora in the alpine zone. New Phytologist 1912;11:37–50.
14. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.
15. Ralescu AL, Ralescu DA. Probability and fuzziness. Information Science 1984;34:85–92.
16. Shapiro DE. The interpretation of diagnostic tests. Statistical Methods in Medical Research 1999;8:113–134. [PubMed: 10501649]
17. Zou KH, Zhou XH. Receiver operating characteristic (ROC) analysis. *Statistics in Epidemiology Report* (*Summer Version*), Statistics in Epidemiology Section, American Statistical Association. URL: http://www.epm.ornl.gov/asasie/newsletter/SIEv2001n1.pdf, 2001.
18. Zhou XH, McClish DK, Obuchowski NA. *Statistical Methods in Diagnostic Medicine.* Wiley: New York, 2002.
19. Zou KH, Hall WJ, Shapiro DE. Smooth nonparametric receiver operating characteristic curves for continuous diagnostic tests. Statistics in Medicine 1997;16:2143–2156. [PubMed: 9330425]
20. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. Journal of Applied Statistics 2000;27:621–631.
21. Zou KH, Hall WJ. Semiparametric and parametric transformation models for comparing diagnostic markers with paired design. Journal of Applied Statistics 2002;29:803–816.
22. Pepe MS. A regression modeling framework for ROC curves in medical diagnostic testing. Biometrika 1997;84:595–608.
23. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics 1998;54:124–135. [PubMed: 9544511]
24. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. Biometrics 2000;56:352–359. [PubMed: 10877289]
25. O'Malley AJ, Zou KH, Fielding JR, Tempany CMC. Bayesian regression methodology for receiver operating characteristic curve with two radiologic applications: prostate biopsy and spiral CT of ureteral stones. Academic Radiology 2001;8:713–725. [PubMed: 11508750]
26. Haralick R, Shanmugan K, Dinstein I. Texture for image classification. IEEE Transactions on Systems, Mechanics, and Cybernetics 1973;3:610–621.
27. Wieand S, Gail MH, James BR, James BR. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 1989;76:585–592.
28. Hettmansperger TP. *Statistical Inference Based on Ranks* 1991, 132–177.
29. Brown D, Rothery P. *Models in Biology: Mathematics, Statistics and Computing* Wiley: New York, 1993.
30. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Statistics in Medicine 1994;13:499–508. [PubMed: 8023031]
31. Chan I, Wells WM, Mulkern RV, Haker S, Zhang J, Zou KH, Maier SE, Tempany CMC. Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted MR imaging; a multi-channel statistical classifier. *Medical Physics* 2003, 2390–2398.
32. Halpern EJ, Albert M, Krieger AM, Metz CE, Maidment AD. Comparison of receiver operating characteristic curves on the basis of optimal operating point. Academic Radiology 1996;3:245–253. [PubMed: 8796672]
33. Sainfort F. Evaluation of medical technologies: a generalized ROC analysis. Medical Decision Making 1991;11 :208–220. [PubMed: 1908934]
34. Englang WL. An exponential model used for optimal threshold selection on ROC curves. Medical Decision Making 1988;8:120–131. [PubMed: 3362039]

35. Shaffer H. Constructing a cut-off point for a quantitative diagnostic test. Statistics in Medicine 1989;8:1381–1391. [PubMed: 2692111]
36. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. Medical Decision Making 1988;8:278–289.
37. Somoza E, Mossman D. Comparing and optimizing diagnostic tests: an information-theory approach. Medical Decision Making 1992;12:179–188. [PubMed: 1513208]
38. Swets JA. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics* Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, 1996, 158–160.
39. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. Radiology 1975;116:533–538. [PubMed: 1153755]
40. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Medical Physics 1996;23:1709–1725. [PubMed: 8946368]
41. Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston JJ, AC. A unified statistical approach for determining significant signals in images of cerebral activation. Human Brain Mapping 1996;4:58–73.
42. Friston KJ, Ashburner J, Poline JB, Frith CD, Heather JD, Frackowiak RSJ. Spatial registration and normalisation of images. Human Brain Mapping 1995;2:165–189.
43. Rajapakse JC, Piyaratna J. Bayesian approach to segmentation of statistical parametric maps. IEEE Transactions on Biomedical Engineering 2001;48:1186–1194. [PubMed: 11585043]
44. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. IEEE Transactions on Medical Imaging 2001;20:45–57. [PubMed: 11293691]
45. Zhang J. The mean field theory in EM procedure for Markov random fields. IEEE Transactions on Signal Processing 1992;40:2570–2583.
46. Qu Y, Hadgu A. A model for evaluating sensitivity and specificity for correlated diagnostics tests in efficacy studies with an imperfect reference test. Journal of American Statistical Association 1998;93:920–928.
47. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology 1995;141:263–272. [PubMed: 7840100]
48. Warfield SK, Zou KH, Kaus MR, Wells WM III. Simultaneous validation of image segmentation and assessment of expert quality. *Proceedings of ISBI Conference* IEEE: New York, vol. 1494, 2002; 1–4.
49. Warfield SK, Zou KH, Wells WM III. Validation of image segmentation and expert quality with an expectation-maximization algorithm. *Proceedings of 5th International Conference on Medical Imaging Computing and Computer Assisted Interventions*, Tokyo, Japan Springer: New York, vol. 2488, 2002; 298–306.
50. Dempster AP, Laird NM, Rubin DB. Validation of image segmentation and expert quality with an expectation-maximization algorithm. Journal of Royal Statistical Society (Series B) 1977;39:34–37.
51. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions.* Wiley: New York, 1997.
52. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Chapman & Hall: London, 1995, 28–32.
53. Hahn GJ, Shapiro SS. *Statistical Models in Engineering.* Wiley: New York. 1994, 95.
54. Cover TMM, Thomas JA. *Elements of Information Theory.* Wiley: New York, 1991.
55. Zou KH, Wells M III, Kaus MR, Kikinis R, Jolesz FA, Warfield SK. Statistical validation of automated probabilistic segmentation against composite latent expert ground truth in MR imaging of brain tumours. *Proceedings of 5th International Conference on Medical Imaging Computing and Computer Assisted Interventions*, Tokyo, Japan Springer: New York, vol. 2488, 2002; 315–322.
56. Greig DM, Porteous BT, Seheult AH. Exact maximum a posterior estimation for binary images. Journal of Royal Statistical Society (Series B) 1989;51:271–279.
57. Kapur T, Grimson L, Wells WM, Kikinis R. Segmentation of brain tissue from magnetic resonance images. Medical Image Analysis 1996;1:109–127. [PubMed: 9873924]

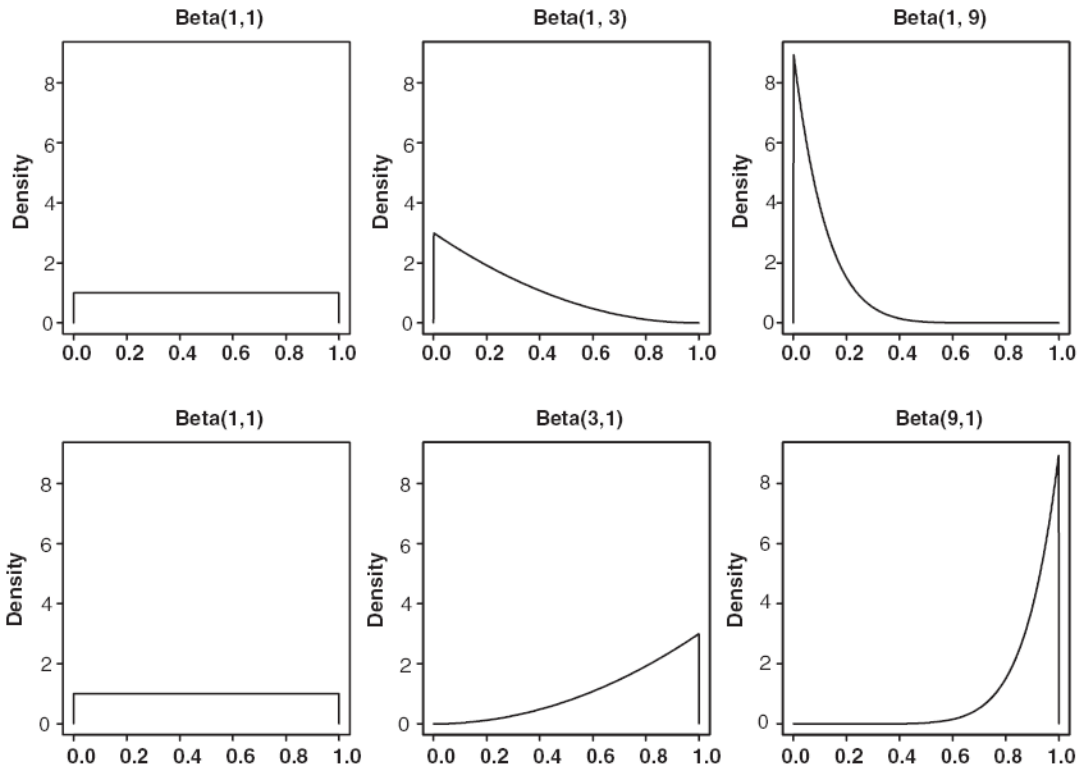

Figure 1.
Beta distributions used in the simulation study for non-tumour (C_0) class only; the Beta distributions for the tumour (C_1) class may be graphed similarly and are omitted here.

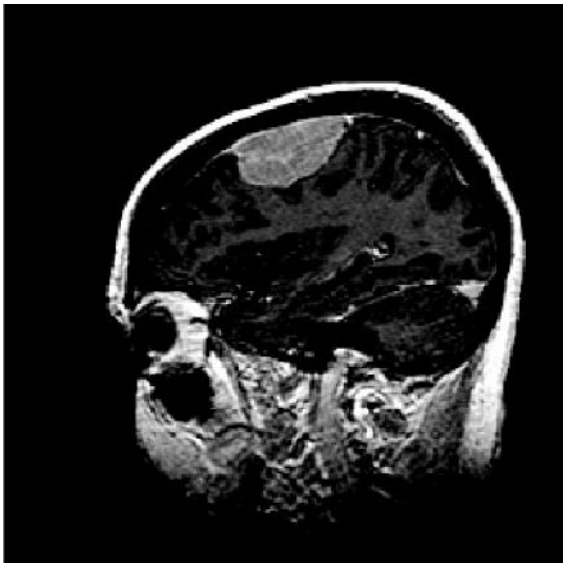

Figure 2.
The grey scale MR image of a case of a meningioma (Case 1).

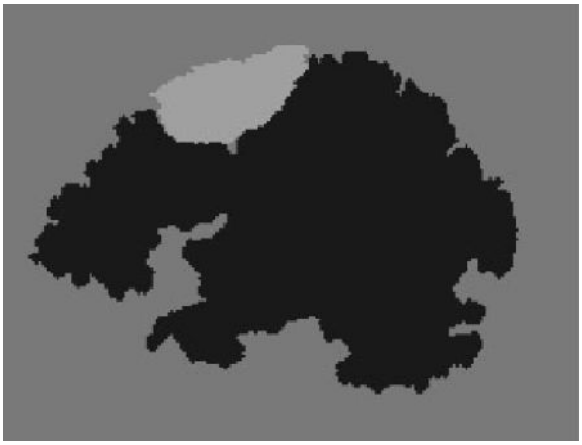

Figure 3.
Estimated composite pixel-wise gold standard of an MRI case of a meningioma (Case 1).

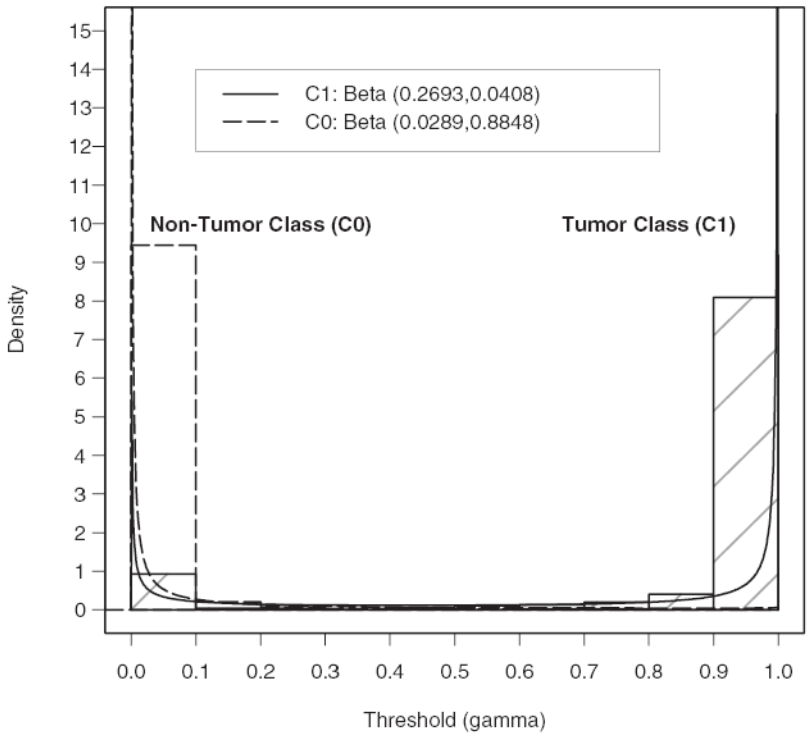

Figure 4.
Estimated mixture of two beta distributions of an MRI case of a meningioma (Case 1).

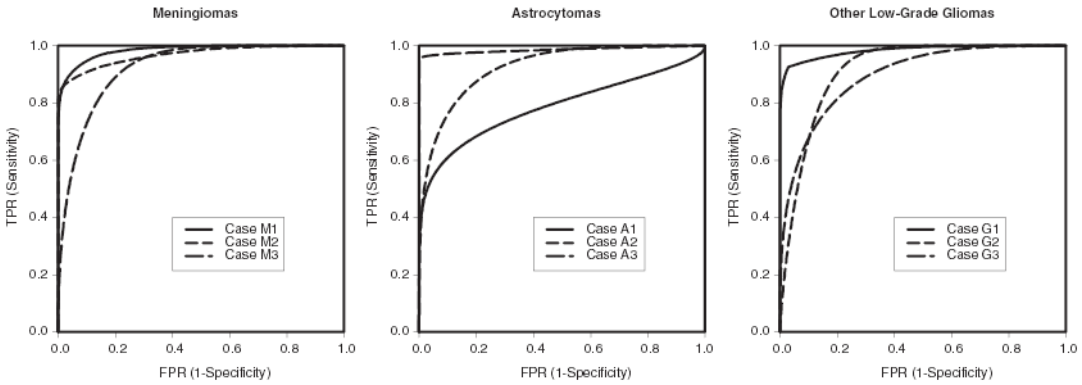

Figure 5.
Estimated ROC curves for the nine brain tumour cases by tumour type: meningiomas (left panel), astrocytomas (centre panel), and other low-grade gliomas (right panel).

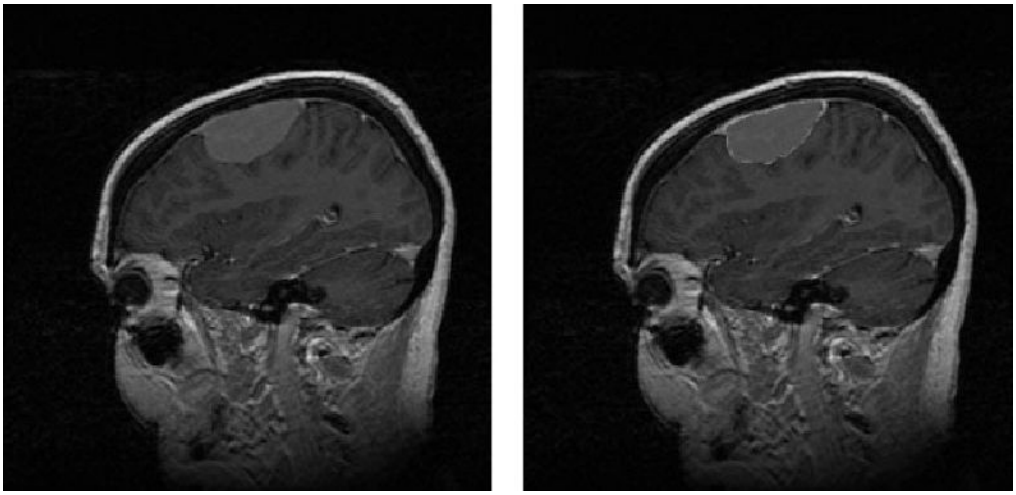

Figure 6.
Pixel-wise frequency of segmentation decision by three expert raters and the estimated pixel-wise gold standard by MRF modelling of a meningioma (Case 1): Left Panel: Frequency of selection; Right Panel: Estimated gold standard by an MRF model, which is virtually identical to that estimated by a pixel-wise independent model.

Table I

A 2 × 2 table of the joint probabilities of the truth (T) vs the corresponding segmentation decision (D_γ) at each possible threshold γ.

Decision vs truth	***T* = 0 (non-tumour)**	***T* = 1 (tumour)**
D_γ = 0 (non-tumour)	p_{11}	p_{12}
D_γ = 1 (tumour)	p_{21}	p_{22}
Marginal total	π	$\bar{\pi}$

where

$p_{11} = P(D_\gamma = 0, T = 0) = P(Z \leq \gamma | T = 0)P(T = 0) = \pi F(\gamma) = \pi p^{-}\gamma$

$p_{21} = P(D_\gamma = 1, T = 0) = P(Z > \gamma | T = 0)P(T = 0) = \pi F^{-}(\gamma) = \pi p_\gamma$

$p_{12} = P(D_\gamma = 0, T = 1) = P(Z \leq \gamma | T = 1)P(T = 1) = \pi^{-} G(\gamma) = \pi^{-} q^{-}\gamma$

$p_{22} = P(D_\gamma = 1, T = 1) = P(Z > \gamma | T = 1)P(T = 1) = \pi^{-} G^{-}(\gamma) = \pi^{-} q\gamma$

Note that the marginal totals, $p_{11} + p_{21} = \pi$ and p12+p22=π^{-}, are related to the Bernoulli parameter of T. Let pγ=$F^{-}(\gamma)$ and qγ=$G^{-}(\gamma)$ (see Section 4.1) [5,6].

Table II

Estimated accuracy metrics (ROC, MI and DSC) and standard errors by Monte Carlo simulations (based on 500 repetitions) under the equal variance assumption, with the number of tumour pixels (n) fixed while the background number of non-tumour pixels (m) varied.

Pixel counts			Specified parameters				Estimated parameters				Estimated metrics		
m	n	$\bar{\pi}$ (per cent)	α_0	β_0	α_1	β_1	$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\widehat{AUC}$	$\widehat{MI}$	$\widehat{DSC}$
1000	1000	50	1	1	1	1	1.003 (0.046)	1.003 (0.048)	1.000 (0.046)	1.000 (0.043)	0.500 (0.013)	0.001 (0.001)	0.451 (0.008)
3000		25					0.999 (0.027)	0.999 (0.027)	1.002 (0.045)	1.001 (0.046)	0.500 (0.011)	0.000 (0.000)	0.299 (0.006)
9000		10					0.999 (0.016)	1.000 (0.015)	1.002 (0.045)	1.000 (0.045)	0.501 (0.010)	0.000 (0.000)	0.152 (0.004)
1000	1000	50	1	1.5	1.5	1	1.002 (0.045)	1.502 (0.069)	1.505 (0.073)	1.003 (0.046)	0.705 (0.011)	0.101 (0.011)	0.554 (0.007)
3000		25					1.003 (0.027)	1.503 (0.041)	1.506 (0.072)	1.003 (0.047)	0.705 (0.009)	0.077 (0.007)	0.402 (0.007)
9000		10					1.000 (0.015)	1.499 (0.024)	1.503 (0.073)	1.001 (0.046)	0.706 (0.009)	0.038 (0.003)	0.234 (0.005)
1000	1000	50	1	2	2	1	1.002 (0.047)	2.005 (0.099)	2.006 (0.095)	1.003 (0.046)	0.833 (0.009)	0.279 (0.017)	0.627 (0.007)
3000		25					0.998 (0.026)	1.998 (0.054)	2.011 (0.095)	1.007 (0.044)	0.833 (0.007)	0.217 (0.010)	0.486 (0.006)
9000		10					1.001 (0.016)	2.002 (0.032)	2.006 (0.094)	1.000 (0.044)	0.834 (0.007)	0.113 (0.005)	0.316 (0.006)
1000	1000	50	1	2.5	2.5	1	1.002 (0.049)	2.510 (0.129)	2.512 (0.128)	1.005 (0.048)	0.908 (0.006)	0.450 (0.018)	0.680 (0.006)
3000		25					1.001 (0.026)	2.502 (0.070)	2.495 (0.128)	0.996 (0.048)	0.908 (0.005)	0.354 (0.011)	0.554 (0.006)
9000		10					1.002 (0.015)	2.505 (0.040)	2.505 (0.121)	1.004 (0.045)	0.908 (0.005)	0.189 (0.006)	0.389 (0.005)
1000	1000	50	1	3	3	1	1.003 (0.045)	3.007 (0.155)	3.016 (0.145)	1.003 (0.044)	0.950 (0.004)	0.589 (0.017)	0.721 (0.005)
3000		25					1.001 (0.027)	3.006 (0.089)	3.000 (0.154)	1.001 (0.048)	0.950 (0.003)	0.468 (0.011)	0.607 (0.006)
9000		10					0.999 (0.015)	2.999 (0.051)	3.019 (0.153)	1.006 (0.048)	0.950 (0.003)	0.256 (0.005)	0.453 (0.005)
1000	1000	50	1	9	9	1	1.003 (0.054)	9.012 (0.512)	9.031 (0.524)	1.004 (0.052)	1.000 (0.000)	0.992 (0.001)	0.889 (0.003)
3000		25					1.003 (0.030)	9.037 (0.292)	9.032 (0.516)	1.003 (0.052)	1.000 (0.000)	0.805 (0.001)	0.840 (0.003)
9000		10					1.002 (0.018)	9.019 (0.183)	8.981 (0.531)	0.998 (0.055)	1.000 (0.000)	0.465 (0.001)	0.768 (0.003)

Table III

Estimated accuracy metrics (ROC, MI and DSC) and standard errors by Monte Carlo simulations (based on 500 repetitions) under the unequal variance assumption, with the number of tumour pixels (n) fixed while the background number of non-tumour pixels (m) varied.

Pixel counts			Specified parameters				Estimated parameters				Estimated metrics		
m	n	$\bar{\pi}$ (per cent)	α_0	β_0	α_1	β_1	$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\widehat{AUC}$	$\widehat{MI}$	$\widehat{DSC}$
1000	1000	50	1	3	1.5	1	0.998 (0.047)	2.993 (0.150)	1.501 (0.070)	1.002 (0.045)	0.847 (0.008)	0.312 (0.016)	0.603 (0.008)
3000		25					1.001 (0.028)	3.003 (0.086)	1.504 (0.068)	1.001 (0.044)	0.848 (0.007)	0.257 (0.011)	0.494 (0.008)
9000		10					1.001 (0.015)	3.001 (0.050)	1.501 (0.071)	1.003 (0.046)	0.847 (0.007)	0.140 (0.006)	0.351 (0.007)
1000	1000	50	1	1.5	3	1	1.001 (0.046)	1.505 (0.075)	3.013 (0.156)	1.006 (0.049)	0.848 (0.009)	0.313 (0.017)	0.667 (0.005)
3000		25					1.000 (0.027)	1.500 (0.041)	3.000 (0.155)	1.001 (0.048)	0.847 (0.006)	0.232 (0.010)	0.503 (0.005)
9000		10					1.000 (0.015)	1.499 (0.023)	3.004 (0.156)	1.002 (0.049)	0.847 (0.006)	0.115 (0.005)	0.308 (0.004)
1000	1000	50	1	9	3	1	0.999 (0.052)	8.988 (0.526)	2.998 (0.155)	1.000 (0.049)	0.995 (0.001)	0.880 (0.010)	0.778 (0.005)
3000		25					1.002 (0.030)	9.018 (0.311)	2.999 (0.146)	1.000 (0.046)	0.995 (0.001)	0.717 (0.006)	0.729 (0.005)
9000		10					1.000 (0.018)	9.000 (0.178)	3.001 (0.156)	1.000 (0.046)	0.995 (0.001)	0.412 (0.004)	0.657 (0.005)
1000	1000	50	1	3	9	1	1.005 (0.050)	3.017 (0.161)	9.015 (0.523)	1.001 (0.051)	0.995 (0.001)	0.882 (0.011)	0.831 (0.003)
3000		25					0.999 (0.027)	3.000 (0.089)	9.003 (0.522)	1.001 (0.052)	0.995 (0.001)	0.704 (0.006)	0.716 (0.003)
9000		10					1.001 (0.016)	3.003 (0.048)	9.011 (0.544)	1.002 (0.056)	0.995 (0.000)	0.396 (0.003)	0.559 (0.003)

Table IV

Estimated accuracy metrics (ROC, MI and DSC) and standard errors by Monte Carlo simulations (based on 500 repetitions) under the equal variance assumption, with the total number of pixels ($N = m + n$) fixed while the proportion of the tumour pixels (n) varied.

Pixel counts			Specified parameters				Estimated parameters				Estimated metrics		
m	n	$\bar{\pi}$ (per cent)	α_0	β_0	α_1	β_1	$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\widehat{AUC}$	$\widehat{MI}$	$\widehat{DSC}$
8500	1500	15	1	1	1	1	1.001 (0.017)	1.001 (0.016)	1.002 (0.037)	1.002 (0.038)	0.500 (0.008)	0.000 (0.000)	0.208 (0.004)
9000	1000	10					0.999 (0.014)	0.999 (0.015)	1.003 (0.047)	1.001 (0.048)	0.500 (0.010)	0.000 (0.000)	0.152 (0.004)
9500	500	5					0.999 (0.014)	1.000 (0.015)	1.002 (0.063)	1.000 (0.062)	0.501 (0.013)	0.000 (0.000)	0.152 (0.003)
8500	1500	15	1	1.5	1.5	1	1.000 (0.016)	1.502 (0.026)	1.499 (0.057)	0.998 (0.036)	0.706 (0.007)	0.054 (0.004)	0.302 (0.005)
9000	1000	10					1.000 (0.015)	1.500 (0.024)	1.497 (0.069)	0.996 (0.043)	0.706 (0.009)	0.039 (0.003)	0.234 (0.005)
9500	500	5					1.000 (0.014)	1.501 (0.022)	1.507 (0.102)	1.003 (0.062)	0.706 (0.011)	0.021 (0.002)	0.143 (0.004)
8500	1500	15	1	2	2	1	0.999 (0.017)	2.001 (0.032)	2.004 (0.076)	1.000 (0.037)	0.834 (0.005)	0.155 (0.006)	0.388 (0.005)
9000	1000	10					1.001 (0.016)	2.001 (0.034)	2.002 (0.097)	1.000 (0.049)	0.833 (0.006)	0.112 (0.005)	0.315 (0.005)
9500	500	5					1.001 (0.015)	2.002 (0.033)	2.004 (0.132)	1.001 (0.065)	0.833 (0.009)	0.062 (0.004)	0.213 (0.006)
8500	1500	15	1	2.5	2.5	1	1.000 (0.016)	2.499 (0.042)	2.497 (0.097)	1.000 (0.037)	0.908 (0.004)	0.255 (0.007)	0.460 (0.005)
9000	1000	10					0.999 (0.016)	2.498 (0.043)	2.510 (0.118)	1.003 (0.045)	0.908 (0.004)	0.189 (0.005)	0.390 (0.005)
9500	500	5					1.001 (0.015)	2.500 (0.040)	2.510 (0.169)	1.003 (0.067)	0.908 (0.006)	0.107 (0.004)	0.283 (0.007)
8500	1500	15	1	3	3	1	1.000 (0.016)	3.000 (0.053)	3.003 (0.119)	1.001 (0.036)	0.950 (0.003)	0.342 (0.007)	0.520 (0.005)
9000	1000	10					1.001 (0.016)	3.004 (0.052)	3.000 (0.123)	1.002 (0.037)	0.950 (0.003)	0.330 (0.006)	0.512 (0.005)
9500	500	10					1.001 (0.015)	3.000 (0.050)	3.013 (0.214)	1.003 (0.067)	0.950 (0.004)	0.148 (0.005)	0.349 (0.007)
8500	1500	15	1	9	9	1	1.000 (0.018)	8.997 (0.183)	8.999 (0.433)	1.000 (0.044)	1.000 (0.000)	0.604 (0.001)	0.800 (0.002)
9000	1000	10					1.002 (0.018)	9.023 (0.183)	9.054 (0.536)	1.004 (0.053)	1.000 (0.000)	0.465 (0.001)	0.786 (0.003)
9500	500	5					1.001 (0.017)	9.010 (0.171)	9.059 (0.761)	1.006 (0.079)	1.000 (0.000)	0.283 (0.001)	0.712 (0.004)

Table V

Estimated accuracy metrics (ROC, MI and DSC) and standard errors by Monte Carlo simulations (based on 500 repetitions) under the unequal variance approach, with the total number of pixels ($N = m + n$) fixed while the proportion of the tumour pixels (n) varied.

Pixel counts			Specified parameters				Estimated parameters				Estimated metrics		
M	n	$\bar{\pi}$ (per cent)	α_0	β_0	α_1	β_1	$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\widehat{AUC}$	$\widehat{MI}$	$\widehat{DSC}$
8500	1500	15	1	3	1.5	1	0.999 (0.016)	3.000 (0.053)	1.501 (0.058)	1.002 (0.037)	0.847 (0.006)	0.188 (0.006)	0.412 (0.006)
9000	1000	10					1.001 (0.017)	3.003 (0.054)	1.504 (0.069)	1.002 (0.046)	0.848 (0.007)	0.140 (0.006)	0.352 (0.007)
9500	500	5					1.001 (0.015)	3.001 (0.050)	1.506 (0.102)	1.007 (0.064)	0.847 (0.010)	0.080 (0.005)	0.259 (0.008)
8500	1500	15	1	1.5	3	1	1.001 (0.016)	1.501 (0.025)	3.004 (0.130)	1.000 (0.040)	0.848 (0.005)	0.161 (0.006)	0.389 (0.004)
9000	1000	10					1.001 (0.015)	1.501 (0.023)	3.007 (0.156)	1.000 (0.047)	0.848 (0.006)	0.116 (0.005)	0.309 (0.004)
9500	500	5					1.000 (0.015)	1.499 (0.023)	3.026 (0.222)	1.007 (0.070)	0.848 (0.008)	0.062 (0.003)	0.197 (0.005)
8500	1500	15	1	9	3	1	1.002 (0.019)	9.012 (0.186)	3.005 (0.126)	0.999 (0.039)	0.995 (0.001)	0.538 (0.004)	0.690 (0.004)
9000	1000	10					1.001 (0.018)	9.011 (0.171)	3.007 (0.158)	1.005 (0.049)	0.995 (0.001)	0.412 (0.004)	0.657 (0.005)
9500	500	5					0.999 (0.017)	9.000 (0.166)	3.003 (0.203)	1.002 (0.065)	0.995 (0.001)	0.250 (0.003)	0.601 (0.007)
8500	1500	15	1	3	9	1	1.000 (0.016)	3.000 (0.051)	9.046 (0.438)	1.004 (0.044)	0.995 (0.000)	0.396 (0.004)	0.560 (0.003)
9000	1000	10					1.000 (0.017)	3.000 (0.053)	9.046 (0.563)	1.004 (0.059)	0.995 (0.000)	0.396 (0.004)	0.560 (0.003)
9500	500	5					1.001 (0.015)	3.001 (0.051)	9.055 (0.789)	1.006 (0.079)	0.995 (0.001)	0.235 (0.003)	0.450 (0.004)

Table VI

Sample statistics and estimated beta parameters for 9 brain tumour cases.

Tumor type	Pixel counts			Sample means and SDs				Estimated beta parameters			
	m	N	$\bar{\pi} = n / N$ (per cent)	$\bar{x}$	s_x	$\bar{y}$	s_y	$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$
M	10534	1175	10	0.0316	0.1264	0.8683	0.2954	0.0289	0.8848	0.2693	0.0408
	15363	1503	8.9	0.0207	0.0890	0.8479	0.3344	0.0321	1.5227	0.1301	0.0233
	12891	1045	7.5	0.1797	0.2746	0.7775	0.2619	0.1716	0.7832	1.1835	0.3387
A	10237	268	2.6	0.3682	0.1548	0.6347	0.2703	3.2081	5.5044	1.3790	0.7937
	11579	1428	11.0	0.1812	0.2496	0.7684	0.2773	0.2500	1.1303	1.0098	0.3043
	7148	1379	16.2	0.0621	0.1229	0.9613	0.1742	0.1773	2.6790	0.2173	0.0087
G	8952	1417	13.7	0.0112	0.0908	0.8693	0.3177	0.0038	0.3394	0.1090	0.0164
	12679	1177	8.5	0.1564	0.2803	0.7398	0.2731	0.1063	0.5732	1.1691	0.4112
	9635	1873	16.3	0.2275	0.2630	0.7369	0.2765	0.3505	1.1903	1.1314	0.4040

Note: M = Meningiomas; A = Astrocytomas; G = Other Low-Grade Gliomas.

Table VII

Estimated accuracy metrics (ROC, MI and DSC) and optimal thresholds for 9 brain tumour cases.

Tumor type	Validation metrics			Optimal thresholds					
	$\widehat{AUC}$	$\widehat{MI}$	$\widehat{DSC}$	$\sqrt{(1-\hat{p})^2+\hat{q}^2}$	$\hat{\gamma}_{opt}$	$\widehat{MI}$	$\hat{\gamma}_{opt}$	$\widehat{DSC}$	$\hat{\gamma}_{opt}$
M	0.9842	0.2888	0.8154	1.3255	0.4709	0.3107	0.8625	0.8730	0.8734
	0.9684	0.3012	0.8415	1.2834	0.8448	0.3065	0.8521	0.8931	0.8268
	0.9242	0.1572	0.4220	1.1844	0.2622	0.1098	0.4657	0.5185	0.8414
A	0.7860	0.0557	0.1970	1.0050	0.7713	0.0415	0.7728	0.4871	0.7808
	0.9255	0.2319	0.5146	1.1881	0.4469	0.1598	0.6843	0.6321	0.8005
	0.9858	0.4649	0.8708	1.4142	1.0000	0.5669	0.8553	0.9724	0.8385
G	0.9829	0.4018	0.8961	1.3720	0.0120	0.4032	0.4905	0.8992	0.6736
	0.9157	0.1595	0.4396	1.1735	0.0773	0.1276	0.2232	0.4897	0.6511
	0.8956	0.2505	0.5276	1.1417	0.4547	0.1693	0.6191	0.6197	0.7113

Note: M = Meningiomas; A = Astrocytomas; G = Other Low-Grade Gliomas.