Journal of Magnetic Resonance 201 (2009) 61–71

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

# Classification of degraded cartilage through multiparametric MRI analysis

Ping-Chang Lin, David A. Reiter, Richard G. Spencer *

*Magnetic Resonance Imaging and Spectroscopy Section, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA*

ARTICLE INFO

*Article history:*
Received 13 May 2009
Revised 7 August 2009
Available online 14 August 2009

*Keywords:*
Multiparametric
Classification
Clustering
Osteoarthritis
Sensitivity
Specificity

ABSTRACT

MRI analysis of cartilage matrix may play an important role in early detection and development of therapeutic protocols for degenerative joint disease. Correlations between MRI parameters and matrix integrity have been established in many studies, but the substantial overlap in values observed for normal and for degraded cartilage greatly limits the specificity of these analyses. We implemented established multiparametric analysis methods to define data clusters corresponding to control and degraded bovine nasal cartilage in two-, three-, and four-dimensional parameter spaces, and applied these results to discriminant analysis of a validation data set. Analyses were performed using the parameters ($T_1$, $T_2$, $k_m$, ADC), where $k_m$ is the magnetization transfer rate and ADC is the apparent diffusion coefficient. Results were compared to univariate analyses. Multiparametric $k$-means clustering led to no improvement over univariate analyses, with a maximum sensitivity and specificity in the range of 60–70% for the detection of degradation using $T_1$, and in the range of 80% sensitivity but only 36% specificity using the parameter pair ($T_1$, $k_m$). In contrast, model-based analysis using more general Gaussian clusters resulted in markedly improved classification, with sensitivity and specificity reaching levels of 80–90% using the pair ($T_1$, $k_m$). Finally, a fuzzy clustering technique was implemented which may be still more appropriate to the continuum of degradation seen in degenerative cartilage disease. In view of its success in identifying mild cartilage degradation, the formal multiparametric approach implemented here may be applicable to the nondestructive evaluation of other biomaterials using MRI.

Published by Elsevier Inc.

## 1. Introduction

Cartilage degeneration, as seen in osteoarthritis, and the attempt to develop therapeutics for degenerative cartilage disease, are important examples of the use of MRI to characterize biomaterials. Indeed, the development of noninvasive MRI techniques for the detection of early osteoarthritis and for monitoring therapeutic response to interventions has been the subject of intense activity over recent years, with, however, limited success [1–3]. While changes in $T_1$, $T_2$, $T_{1\rho}$, magnetization transfer ratio (MTR) and rate ($k_m$), apparent diffusion coefficient (ADC), and $T_1$ in the presence of gadolinium (dGEMRIC) have been observed to accompany cartilage degradation, with, for example, MTR and $k_m$ exhibiting somewhat enhanced specificity to changes in the collagen component and the dGEMRIC index being preferentially specific to proteoglycan [3–8], these MRI parameters exhibit limited sensitivity to cartilage pathology and limited specificity for particular cartilage matrix components. Even when a statistically significant difference in the arithmetic mean of a given MRI parameter is observed between control and pathologic cartilage [9–11], there remains a substantial degree of overlap in parameter values between groups.

Therefore, cartilage degeneration represents an ideal setting for the application of multiparametric MRI evaluation of biomaterials. We hypothesized that formal multiparametric analysis would improve the sensitivity and specificity of MRI evaluation of cartilage as compared to univariate analysis. We implemented two analytic approaches to the classification of pathomimetically-degraded cartilage samples, discriminant analyses based on k-means clustering and model-based clustering. $k$-Means clustering is based on the attempted partition of a training set into disjoint subgroups about distinct centroids. We took $k = 2$, indicating the presence of clusters that correspond to control or to degraded cartilage, while the number of MRI outcome measures incorporated into the characterization of the data defines $N$. Classification of a validation set sample is then determined by assignment to whichever of the two clusters has the smaller $N$-dimensional Euclidean distance between its corresponding centroid and the sample's coordinates in normalized parameter space. In model-based cluster analysis, $k$ and $N$ are defined in the same fashion, and a parameterized model is developed to optimize the fitting of the sample distribution within a training set. The model is then applied to a validation set [12]. Sensitivities and specificities of these multiparametric

* Corresponding author. Address: Magnetic Resonance Imaging and Spectroscopy Section, NIH/National Institute on Aging, GRC 4D-06, Baltimore, MD 21224, USA. Fax: +1 410 558 8318.
*E-mail address:* spencer@helix.nih.gov (R.G. Spencer).

1090-7807/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jmr.2009.08.001

analyses were compared with those of univariate analysis. In order to render the results potentially applicable to the detection of early cartilage degeneration, analysis was performed on samples subjected to two mild enzymatic digestion protocols, using trypsin and collagenase, respectively. Further, to limit the confounding influence of spatial heterogeneity of the samples at this stage of the development of the methodology, digestion was applied to samples of bovine nasal cartilage (BNC). Finally, to demonstrate the applicability of the model-based clustering approach to ranges of cartilage degradation, such as is seen in OA, we implemented a procedure based on probabilistic assignments to clusters.

## 2. Materials and methods

### 2.1. Sample preparation

Cartilage discs of 8 mm diameter were excised from nasal septa of freshly slaughtered ~6 month-old calves (Green Village Packing, Green Village, NJ). After removal of a central 2.5 mm plug, 5–6 discs were threaded onto each of four hollow tubes, which were then inserted into the wells of a susceptibility-matched four-well sample holder containing DPBS buffer at pH 7.5 ± 0.1. Degradation was performed by sample incubation in DPBS buffer with 1 mg/ml trypsin (6 h; Sigma–Aldrich, St. Louis, MO) or 30 units/ml collagenase type II (16 h; Worthington Biochemical Corp., Lakewood, NJ) at 37 °C in a 5% $CO_2$ atmosphere, representing relatively mild degradations. A total of 40 control, 40 trypsin-digested and 40 collagenase-digested samples were studied. Analysis was performed separately on (i) 80 samples consisting of the controls plus trypsin-degraded samples, and (ii) 80 samples consisting of the controls plus collagenase-degraded samples. Control samples were the same as those used for a previous study [14].

### 2.2. MRI measurements

Data were acquired using a 9.4T/105-mm Bruker DMX spectrometer (Bruker Biospin GmbH, Rheinstetten, Germany) equipped with a 30-mm proton birdcage resonator. Images were taken from 0.5 mm-thick sagittal slices through the center of each of the four wells in the sample holder, permitting all samples within a given well to be imaged simultaneously. Sample temperature was maintained at 4.0 ± 0.1 °C by cold air from a vortex tube (Exair, Cincinnati, OH). $T_2$ data were acquired using a 64-echo CPMG pulse sequence with TR/TE = 5 s/12.8 ms. To perform $T_1$ mapping, a progressive saturation spin-echo sequence with TE = 12.8 ms was employed, where TR was varied from 100 ms to 15 s in 12 steps. MT data were obtained using the same spin-echo sequence preceded by a 6 kHz off-resonance saturation pulse of amplitude $B_1$ = 12 μT and pulse length $t_p$ incremented from 0.1 to 4.6 s in eight steps. Measurements of apparent diffusion coefficient (ADC) were performed with the diffusion-sensitizing gradients oriented along the direction of the $B_0$ field and perpendicular to the sample surface. A spin-echo sequence was used (TR = 5 s), with the diffusion-sensitizing pulses of duration $\delta$ = 5 ms placed on either side of the 180° refocusing pulse, and with a constant interval of $\Delta$ = 13.8 ms between the gradient pulse centers. The diffusion-sensitizing gradient strength was increased from 0 to 320 mT/m in eight steps. Other MRI parameters included BW = 50 kHz, NEX = 2, FOV = 4.0 × 1.5 cm (read × phase encode), matrix size = 256 × 128, and resolution = 156 × 117 μm. The total acquisition time of these four measurements for samples in a four-well sample holder was about 8 h.

Signal intensity was averaged over all pixels in a region of interest (ROI) covering an entire BNC disk. Averaged intensities were then fit to appropriate three-parameter monoexponential functions to obtain $T_1$, $T_2$, MT ratio (MTR), $T_{1sat}$, $k_m$ = MTR/$T_{1sat}$, and ADC [13], which were calculated independently for each sample. The maximum coefficients of variation (CV) of fits of MR data to appropriate functional forms have been reported previously [14], and were 4%, 6%, 16% and 11% for $T_1$, $T_2$, $k_m$, and ADC, respectively. The mean CVs were much smaller than these values. These values are smaller than or on the order of sample heterogeneity with respect to MRI parameters, indicating that fitting errors, due either to finite SNR or potentially incorrect data models, are negligible in our analysis. In terms of data models, we note that multicomponent $T_2$ relaxation can be observed in comparable systems [15]; we did not extend our analysis to incorporate this. In addition, our experiments were not appropriate for delineating multiple components in the $T_1$, $k_m$ and ADC data curves, given the acquisition of only 8 or 12 data points.

### 2.3. Assessment of sensitivity and specificity

The two data sets analyzed (each with $n$ = 80 samples total; $n$ = 40 control and $n$ = 40 degraded) consisted of the control samples plus the trypsin-treated samples, and the same control samples plus the collagenase-treated samples. These two data sets were analyzed entirely independently. The training set was formed from a random selection of 53 out of 80 samples (equal to 2/3 of the total), with the remaining samples serving as the validation set [16]. In order to avoid potential chance selection of a particularly favorable or unfavorable training set, 100 independent realizations of random training set selection were performed for each analysis, and the results averaged and reported. Four classification approaches were studied: (i) mean values, (ii) $k$-means clustering, (iii) model-based discriminant analysis with restriction to single-component clusters, and (iv) full model-based discriminant analysis with multiple-component clusters permitted [17,18]. All analyses were performed using the MCLUST package written in the R language [19], and in-house designed routines written in the MATLAB 7.4 programming language (The MathWorks Inc., Natick, MA).

Sensitivity was defined in the usual way as the rate of true positives, that is, the number of correctly classified degraded samples divided by the total number of degraded samples. Similarly, specificity was defined as the rate of true negatives, that is, the number of correctly classified control samples divided by the total number of controls.

### 2.4. Assignments using arithmetic means of individual MRI parameters

The mean value of a specified MRI parameter was calculated separately for the control and degraded samples in the training set. Each sample in the validation set was then classified as control or degraded, depending upon which of these two means its MRI parameter value was closer to. Although the quality of the model is determined by the classification accuracy within the validation set, the same classification procedure was also applied to the training set itself as an indicator of the maximal ability of arithmetic means to distinguish between the control and degraded samples.

### 2.5. Assignments using multiparametric k-means clustering

In this approach, each sample is assigned an $N$-tuple defined by a set of $N$ measured parameters. The algorithm then partitions the set of data points corresponding to the samples into $k$ disjoint clusters in $N$-dimensional Euclidean space, with each sample ultimately assigned to the cluster for which the Euclidean distance from the sample point to that cluster's centroid is the minimum over all clusters [18]. In our case, $N$ = 2, 3, or 4, corresponding to use of two or more of the measured parameters $T_1$, $T_2$, $k_m$, and

ADC, while $k = 2$, corresponding to clusters being defined for control and for degraded tissue. Cluster definition was performed using the training set data. Each validation set sample was then assigned to the control or the degraded group depending upon which of the two cluster centroids its location in parameter space was closer to. Analysis was performed using the same randomly-selected training and validation sets as were used for the univariate analyses. We used normalized parameter values in order to, in effect, assign equal weights to each measured parameter.

### 2.6. Assignments using model-based cluster analysis with and without restriction to single-component clusters

A limitation of $k$-means clustering as described above is that all clusters are by definition hyperspheres of equal volume; this model however is appropriate only for a very limited set of data distributions. A greater degree of flexibility in cluster morphology is achieved by adopting a model-based algorithm, through which each cluster is defined according to Gaussian components [20]. Each component, $g$, of a cluster is described as an $N$-dimensional ellipsoid centered at mean $\boldsymbol{\mu}_g$ and with an associated covariance matrix $\Sigma_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^{\mathbf{T}}$ [17]. The indicated eigenvalue decomposition of $\boldsymbol{\Sigma}_g$ is written in terms of $\lambda_g$, which sets the volume of component $g$, $\mathbf{D}_g$, a matrix of eigenvectors defining its orientation, and $\mathbf{A}_g$, a diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_g$, defining its shape. This approach was implemented using the MCLUST package, based on the statistical software R [17], which examines several different covariance structures, as defined by the volumes, orientations, and shapes of Gaussian components. The selection of the appropriate model for the training set data is performed through a maximum likelihood approach in which the likelihood function for a particular Gaussian mixture model is calculated from the products of the individual probabilities of the data points belonging to an assigned cluster. The number of parameters in the likelihood function for a Gaussian mixture model $\prod_{i=1}^{n} \sum_{g=1}^{G} \tau_g \varphi_g(\boldsymbol{x}_i; \mu_g, \Sigma_g)$, where $\varphi_g(\boldsymbol{x}_i; \mu_g, \Sigma_g)$ denotes a cluster component described by a multiparametric Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix of the data points assigned to g $\boldsymbol{\Sigma}_g$, and with $\boldsymbol{\tau}_g$ denoting the probability that a multidimensional data point $\boldsymbol{x}_i$ belongs to this component, is selected for the training set according to the Bayesian information criterion (BIC) [21]:

$$\mathrm{BIC} = 2 \times \mathrm{loglikelihood} - \mathrm{number\ of\ model\ parameters} \times \log(\mathrm{number\ of\ samples}).$$

In this formulation, increased closeness of fit through additional parameterization is penalized by model complexity, and the model selected is the one for which the BIC is maximized. After model specification through BIC maximization, validation samples are assigned to a subgroup according to that sample's highest posterior probability [12].

In addition to the analysis described above, a somewhat more restrictive approach was also implemented in which the model-based clusters were restricted to single components. This maintains contact with the $k$-means approach while permitting more flexible cluster morphology.

### 2.7. Statistical analysis

Prior to multiparametric analysis, each MRI parameter set for each treatment group was standardized to zero mean and unit variance through the transformation $p_i \rightarrow (p_i - \mu)/\sigma$ where $p_i$ is a measured value of a given parameter for sample $i$, and $\mu$ and $\sigma$ are the calculated mean and standard deviation of the $\{p_i\}$. For illustrative purposes, error ellipsoids were calculated for cluster components from the covariance matrix, $\Sigma$, of sets of MRI parameter values. These ellipsoids are contours of constant probability density in parameter space, centered at the vector of parameter means $\mu$, with the contour defined as the locus of points $\boldsymbol{c}$ in parameter space that satisfy $(\boldsymbol{c} - \mu)' \Sigma^{-1} (\boldsymbol{c} - \mu) = \chi_p^2(\alpha)$. The directed lengths of the ellipsoid semi-axes are $k\sqrt{\lambda_i}\boldsymbol{e_i}$, where $k$ is the square root of $\chi_p^2(\alpha)$ and $\lambda_i$ and $\boldsymbol{e_i}$ are the eigenvalues and eigenvectors of the covariance matrix. The value of $\chi_p^2(\alpha)$ is calculated from the $\chi^2$ probability distribution with $p$ degrees of freedom such that the contour as defined above contains a proportion $(1 - \alpha)$ of the probability distribution reflected by the cluster component model. We have selected $\alpha = 0.5$ throughout. Finally, averages of MRI parameters from non-degraded and degraded samples were reported as ±SD with $p < 0.05$ evaluated by the two-tailed Student $t$-test considered statistically significant.

### 2.8. Probabilistic assignment to clusters

A fuzzy clustering-based approach may represent a more realistic classification procedure than binary discriminant analysis for tissue samples with differing degrees of degradation. With this approach, each sample is assigned with a certain probability to a control or a degraded group with the sum of these two respective probabilities equaling unity. We implemented this by reporting the conditional probability of assignment to a cluster component resulting from the expectation-maximization algorithm [12]. In order to provide an unambiguous probability assignment, we evaluated this for the case of single-component clusters; in practice, given the similarity in the results shown in Tables 4 and 5, this is not a limitation.

## 3. Results

Fig. 1A shows a transverse slice from a spin-echo image of the four-well sample holder. Slices (1.5 cm wide × 0.5 mm thick × 4 cm in length along the well-axis) used for ROI definitions are indicated by rectangular outlines. Fig. 1B shows ROI selection within such a slice on a diffusion-weighted image of a single well containing five BNC discs.

Table 1 shows MRI parameter means for the three groups studied. As expected, upon digestion with either collagenase or trypsin, mean $T_1$, $T_2$ and ADC increased while $k_m$ decreased [4]. For this mild degradation protocol, there were no statistically significant differences between the control BNC discs and the enzymatically digested BNC samples for any of these MRI parameters, with the exception of $T_1$, for which the groups nevertheless differed by only 10%. These results indicate a limited capability of assigning samples with unknown status into non-degraded and degraded subgroups using any of these single parameters. This is demonstrated explicitly in Table 2, showing the results of classification according to arithmetic means calculated from training sets. Of note is that there were no substantial differences between accuracy in classifying the validation as compared to the training sets for any parameters, for both enzymatic degradation categories. This lack of dependence of classification accuracy on the randomly selected training sets indicates that the simple means model is not highly constrained by the training set. For both enzymatic treatments, classification via $T_1$ resulted in the best results, with sensitivity and specificity above 60%. This is consistent with the fact that $T_1$ was the only parameter showing significant differences between the control and digested BNC samples (Table 1). The other parameters, $T_2$, $k_m$ and ADC, were all comparable to one another in performance. As an example, although $T_2$ and $k_m$ are often considered as being relatively specific markers for collagen content, classification of collagenase-digested vs. control samples according to these parameters did not result in particularly accurate results. Overall, these analyses indicate the limited sensitivities and specificities

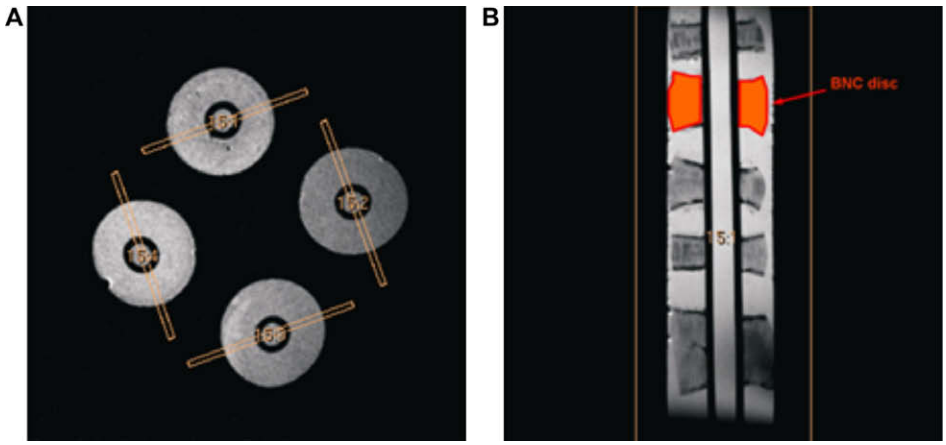


**Fig. 1.** Images of BNC samples within the 4-well sample holder following 6-h trypsin digestion. (A) Transverse slice showing all four wells in cross section. (B) Diffusion-weighted image used for delineation of ROIs, as indicated.

**Table 1**
MRI parameter means in control and degraded BNC samples.

| | $T_1$ (ms) | $T_2$ (ms) | $k_m$ ($s^{-1}$) | ADC ($\times 10^{-4}$ $mm^2/s$) |
|---|---|---|---|---|
| Control ($n = 40$)[a] | 1208 ± 101 | 55.0 ± 11.1 | 0.87 ± 0.22 | 9.29 ± 1.00 |
| 6-h trypsin ($n = 40$) | 1303 ± 93* | 59.0 ± 10.3 | 0.82 ± 0.22 | 9.55 ± 1.02 |
| 16-h collagenase ($n = 40$) | 1335 ± 108* | 60.0 ± 12.0 | 0.82 ± 0.21 | 9.70 ± 0.91 |

* $p < 0.05$ control vs. trypsin-digested or control vs. collagenase-digested.
[a] Ref [14].

**Table 2**
Univariate classification of control and degraded BNC samples.

| Category | Parameter | Training set | | Validation set | |
|---|---|---|---|---|---|
| | | Sensitivity[a] | Specificity[b] | Sensitivity | Specificity |
| Control vs. 6-h trypsin | $T_1$ | 0.67 ± 0.05 | 0.63 ± 0.06 | 0.67 ± 0.12 | 0.62 ± 0.13 |
| | $T_2$ | 0.55 ± 0.05 | 0.55 ± 0.04 | 0.55 ± 0.17 | 0.53 ± 0.14 |
| | $k_m$ | 0.62 ± 0.08 | 0.45 ± 0.08 | 0.59 ± 0.16 | 0.43 ± 0.12 |
| | ADC | 0.59 ± 0.05 | 0.56 ± 0.05 | 0.56 ± 0.13 | 0.56 ± 0.14 |
| Control vs. 16-h collagenase | $T_1$ | 0.67 ± 0.04 | 0.69 ± 0.05 | 0.68 ± 0.11 | 0.68 ± 0.12 |
| | $T_2$ | 0.48 ± 0.05 | 0.57 ± 0.05 | 0.46 ± 0.14 | 0.57 ± 0.15 |
| | $k_m$ | 0.59 ± 0.06 | 0.45 ± 0.07 | 0.55 ± 0.12 | 0.43 ± 0.11 |
| | ADC | 0.6 ± 0.05 | 0.61 ± 0.06 | 0.58 ± 0.12 | 0.61 ± 0.17 |

The training set contains 53 randomly-selected samples while the remaining 27 samples form the validation set in all cases. Values shown are for the average of 100 independent selections of the training set and associated validation set.
[a] Sensitivity is defined as the ratio of correctly assigned degraded samples to the total number of degraded samples.
[b] Specificity is defined as the ratio of correctly assigned control samples to the total number of control samples.

achieved through classification of these mildly-degraded samples using arithmetic means of single MRI parameters.

Fig. 2 shows pairwise displays of the standardized values of $T_1$, $T_2$, $k_m$ and ADC and their corresponding error ellipses for control and collagenase-treated samples. Fig. 2A illustrates the locus of sample points as defined by ($T_1$, $k_m$) pairs. The angle between the semi-major axis of the error ellipse and the horizontal axis was 142° initially and decreased to 138° with degradation (clockwise rotation; CW); this rotation was accompanied by a 29% reduction in area. For ($T_1$, $T_2$) coordinates, the error ellipse retained an angle of 50° after collagenase digestion, while the area increased by 8% (Fig. 2B). Fig. 2C illustrates the corresponding result for ($T_2$, $k_m$) coordinates; the error ellipse rotated from 139° to 133° (CW) and decreased in volume by 11% with digestion. The ($T_1$, ADC) analysis demonstrated an error ellipse rotation by 6° (CW), from 53° to 47°, and a 10% reduction in area (Fig. 2D), while for ($T_2$, ADC) coordinates, the error ellipse rotated from 48° to 41° (CW) and gained 39% in area (Fig. 2E). Fig. 2F shows the analysis using ($k_m$, ADC); in this case, the error ellipse again rotated minimally from 135° to 137° (counterclockwise rotation; CCW), with a loss in area of 23%. Thus, collagenase digestion resulted in minimal rotations of the error ellipses, that is, minimal changes in covariance matrix structure, for a given pair of parameters, but a much larger range of changes in area. Similar results were seen with trypsin digestion (data not shown). Of course, whether error ellipse rotation is CW or CCW merely reflects the display choice, but given that choice, indicates the potential change in steepness of the dependence of one variable on another.

Table 3 shows classification sensitivities and specificities of control vs. trypsin-degraded samples achieved through use of multiparametric *k*-means clustering. As was also the case for the univariate analysis, the accuracy in the validation set was essentially indistinguishable from that in the training set. The sensitivity of classification was greater than that of univariate classification when using the parameter sets ($T_1$, $k_m$) or ($T_1$, $k_m$, ADC). However, simultaneous improvement in sensitivity and specificity as compared to univariate analysis was not achieved for any parameter set. Similar results were obtained for collagenase-degraded samples (data not shown). Thus, analysis through use of simple *k*-means clustering was of limited value in classification.

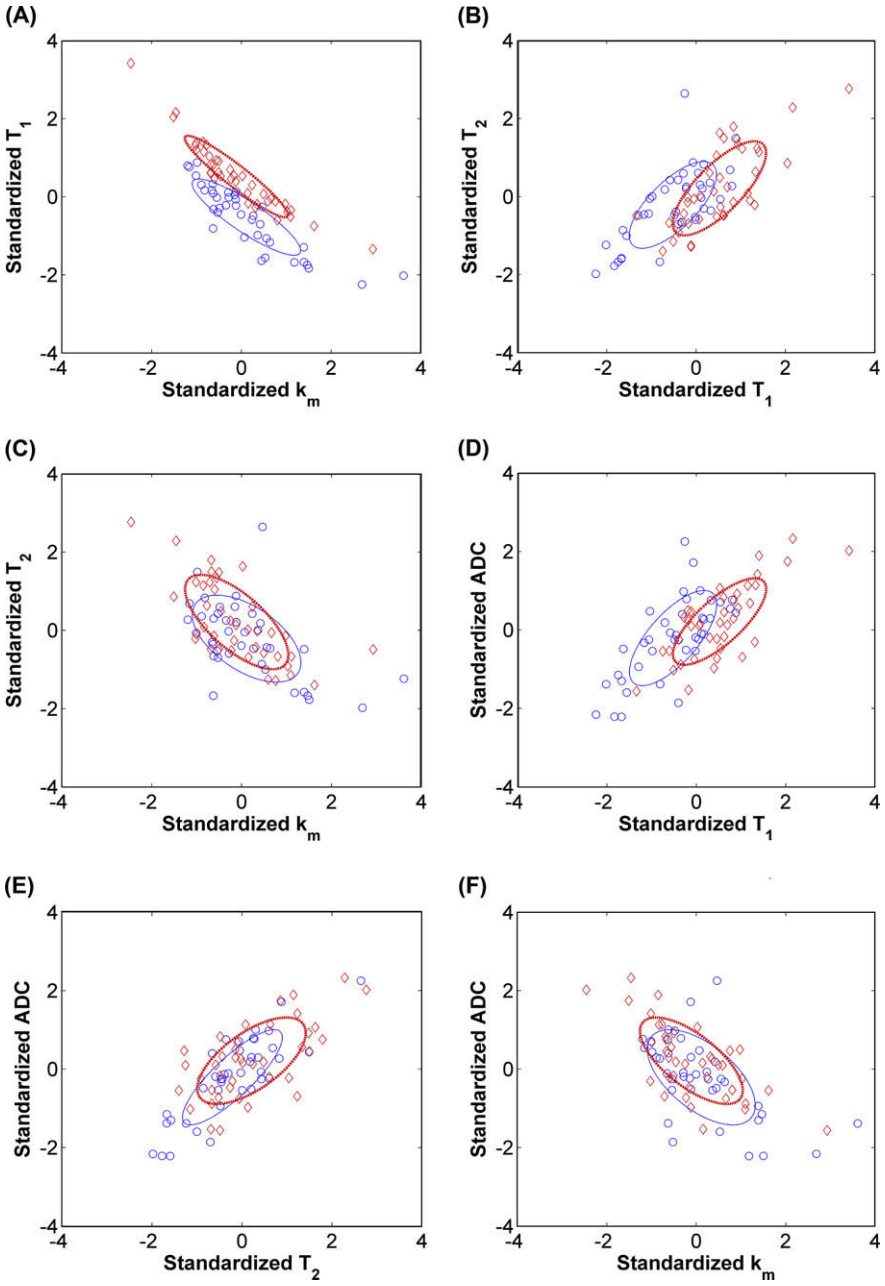


**Fig. 2.** Bivariate scatter plots of data acquired from control and collagenase-digested BNC samples. Covariance matrices for each pair of MRI parameters were calculated for control (blue; O) and degraded (red; ◇) BNC samples. The corresponding error ellipses illustrate the relationship between the parameter pairs indicated. (A) ($k_m$, $T_1$), (B) ($T_1$, $T_2$), (C) ($k_m$, $T_2$), (D) ($T_1$, ADC), (E) ($T_2$, ADC), and (F) ($k_m$, ADC) parameter pairs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 3 shows plots of control and trypsin-degraded samples in 3-dimensional parameter space, and the results of classification of a validation set using model-based discriminant analysis as implemented through MCLUST. The particular example shown is one of the 100 random selections of the training set used to define sensitivity and specificity of the parameter set ($T_1$, $T_2$, $k_m$).

As seen in Fig. 3A, for this particular selection of training set and validation set, restriction to a single Gaussian component per cluster resulted in sensitivity of 96% and specificity of 80% for the training set, with lower values of 75% and 66.7%, respectively, being achieved for the validation set (Fig. 3B). Accuracy was improved when the single-component restriction was removed, with sensitivity of 96.4% and specificity of 96% being achieved in the training set (Fig. 3C) and 91.7% and 80%, respectively, in the validation set (Fig. 3D).

Table 4 is a tabulation of the results of the model-based approach in which clusters were restricted to a single component. Clearly, overall performance was much improved as compared to *k*-means clustering. For both enzymatic degradation protocols, the most accurate classifications were obtained using the parameter combinations ($T_1$, $k_m$), ($T_1$, $T_2$, $k_m$), ($T_1$, $k_m$, ADC) and ($T_1$, $T_2$, $k_m$, ADC) with sensitivity and specificity of approximately 80% or above for the training set and above 75% for the validation set. These results clearly indicate that improved performance in both training set and validation set classification is made possible by a more flexible description of the dataset in parameter space. Of note is that the decrease in classification accuracy for the validation set as compared to the training set was somewhat more pronounced than for classification based on arithmetic means and *k*-means

**Table 3**
Discriminant analysis of control and degraded BNC samples using *k*-means clustering.

| Category | Parameters | Training set | | Validation set | |
|---|---|---|---|---|---|
| | | Sensitivity | Specificity | Sensitivity | Specificity |
| Control vs. 6-h trypsin | ($T_1$, $T_2$) | 0.63 ± 0.07 | 0.54 ± 0.08 | 0.63 ± 0.13 | 0.54 ± 0.13 |
| | ($T_1$, $k_m$) | 0.79 ± 0.07 | 0.38 ± 0.08 | 0.79 ± 0.12 | 0.36 ± 0.11 |
| | ($T_1$, ADC) | 0.7 ± 0.08 | 0.51 ± 0.1 | 0.69 ± 0.11 | 0.49 ± 0.15 |
| | ($T_2$, $k_m$) | 0.52 ± 0.17 | 0.54 ± 0.17 | 0.44 ± 0.19 | 0.48 ± 0.21 |
| | ($T_2$, ADC) | 0.68 ± 0.1 | 0.46 ± 0.09 | 0.63 ± 0.17 | 0.43 ± 0.14 |
| | ($k_m$, ADC) | 0.63 ± 0.2 | 0.44 ± 0.18 | 0.57 ± 0.23 | 0.38 ± 0.21 |
| | ($T_1$, $T_2$, $k_m$) | 0.68 ± 0.1 | 0.44 ± 0.1 | 0.66 ± 0.16 | 0.41 ± 0.12 |
| | ($T_1$, $T_2$, ADC) | 0.68 ± 0.1 | 0.49 ± 0.11 | 0.66 ± 0.15 | 0.47 ± 0.14 |
| | ($T_1$, $k_m$, ADC) | 0.73 ± 0.1 | 0.42 ± 0.1 | 0.71 ± 0.16 | 0.38 ± 0.13 |
| | ($T_2$, $k_m$, ADC) | 0.61 ± 0.16 | 0.46 ± 0.14 | 0.54 ± 0.2 | 0.41 ± 0.14 |
| | ($T_1$, $T_2$, $k_m$, ADC) | 0.67 ± 0.13 | 0.45 ± 0.11 | 0.62 ± 0.18 | 0.42 ± 0.13 |

Analysis was performed using the same training and validation sets as for Table 2. Values shown are for the average of 100 independent selections of the training set and associated validation set.

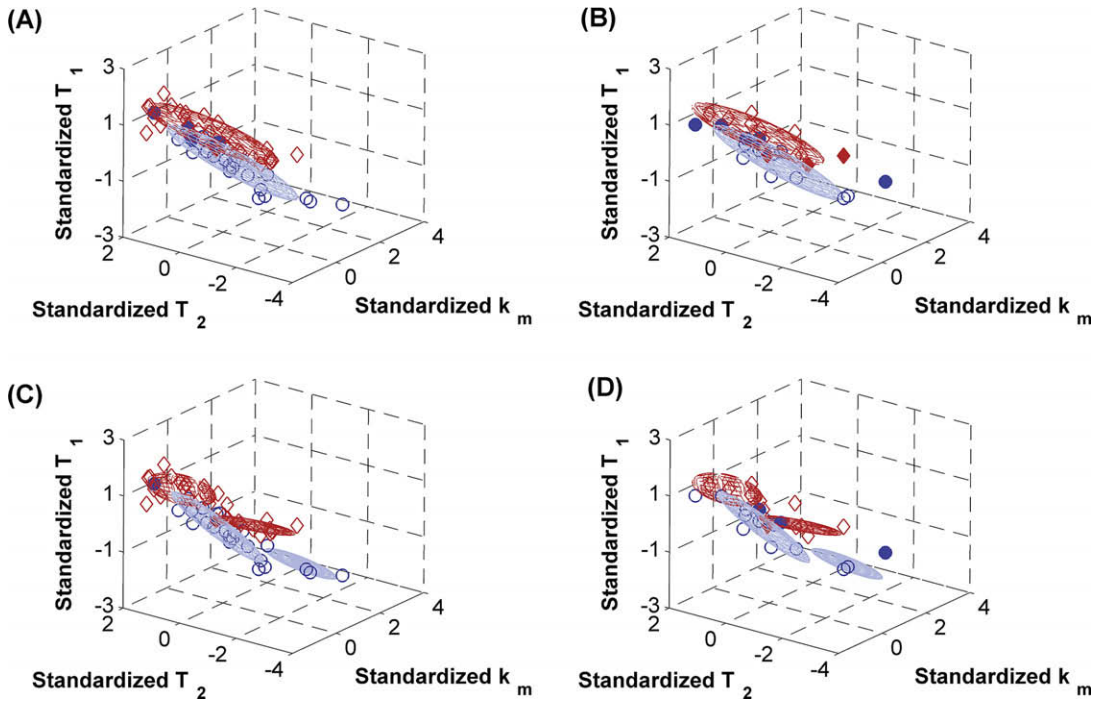


**Fig. 3.** Model-based discriminant analysis based on MCLUST partitioning of control (blue; O) and trypsin-degraded (red; ◇) BNC samples. Models were developed based on 53 samples selected randomly from the total set of 80, with the remaining 27 samples forming the validation set shown. The MRI parameters $T_1$, $T_2$ and $k_m$ are shown here for illustrative purposes. Error ellipsoids represent the 50% confidence surface. Incorrectly classified samples are indicated by solid symbols. Panels A and B illustrate the results with clusters restricted to single components, while Panels C and D illustrate clusters formed from multiple (two, in this case) components. Panel A: training set; five samples in the control group (specificity 80%) and one sample in the trypsin-degraded group (sensitivity 96.4%) were misclassified using single-component clusters. Panel B: validation set, with five control and three degraded samples being misclassified, resulting in sensitivity and specificity of 75% and 67%, respectively, using single-component clusters. Panel C: training set; a single sample in both the control and degraded groups were misclassified, resulting in a sensitivity of 96.4% and a specificity of 96% using multiple-component clusters. Panel D: validation set, with three control samples and one degraded sample being misclassified, resulting in sensitivity and specificity of 91.7% and 80%, respectively, using multiple-component clusters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

clustering. This is consistent with the more highly constrained fitting of the data by the constructed model.

Table 5 shows the results of model-based classification in which the restriction to single-component clusters was removed. As shown, this resulted in further improvement in the sensitivity and specificity of training set classification, but not in classification in the validation set. The parameter combinations ($T_1$, $k_m$), ($T_1$, $T_2$, $k_m$) and ($T_1$, $k_m$, ADC) again performed better than classification according to any single parameter or parameter combination using arithmetic means or *k*-means clustering. As for the analysis with single-component clusters, the sensitivities and specificities in the validation sets were lower than for the training sets. This, together with relatively large standard deviations, indicates that this model-based algorithm was sensitive to the random selection of validation data points via construction of a relatively highly-constrained model.

We performed an additional analysis of all 120 samples ($n$ = 40 control, $n$ = 40 trypsin-degraded, $n$ = 40 collagenase-degraded) taken together, to determine the accuracy of assignment to the control group or the degraded group, with the degraded group now comprised of both of the enzymatically-degraded sets of samples. The training set again consisted of a random selection of 2/3 of the entire set of samples, and the results of 100 independent random assignments to training set and validation set were averaged.

**Table 4**
Discriminant analysis of control and degraded BNC samples using single-component Gaussian clusters.

| Category | Parameters | Training set | | Validation set | |
|---|---|---|---|---|---|
| | | Sensitivity | Specificity | Sensitivity | Specificity |
| Control vs. 6-h trypsin | ($T_1$, $T_2$) | 0.74 ± 0.07 | 0.65 ± 0.06 | 0.68 ± 0.13 | 0.61 ± 0.12 |
| | ($T_1$, $k_m$)* | 0.84 ± 0.04 | 0.79 ± 0.06 | 0.81 ± 0.11 | 0.78 ± 0.11 |
| | ($T_1$, ADC) | 0.68 ± 0.08 | 0.65 ± 0.07 | 0.61 ± 0.13 | 0.59 ± 0.12 |
| | ($T_2$, $k_m$) | 0.61 ± 0.15 | 0.53 ± 0.16 | 0.51 ± 0.16 | 0.43 ± 0.18 |
| | ($T_2$, ADC) | 0.61 ± 0.06 | 0.75 ± 0.07 | 0.60 ± 0.13 | 0.69 ± 0.13 |
| | ($k_m$, ADC) | 0.65 ± 0.16 | 0.59 ± 0.21 | 0.54 ± 0.18 | 0.52 ± 0.20 |
| | ($T_1$, $T_2$, $k_m$)* | 0.86 ± 0.04 | 0.80 ± 0.06 | 0.81 ± 0.11 | 0.75 ± 0.11 |
| | ($T_1$, $T_2$, ADC) | 0.73 ± 0.08 | 0.76 ± 0.06 | 0.67 ± 0.14 | 0.65 ± 0.15 |
| | ($T_1$, $k_m$, ADC)* | 0.86 ± 0.04 | 0.82 ± 0.05 | 0.80 ± 0.11 | 0.77 ± 0.11 |
| | ($T_2$, $k_m$, ADC) | 0.64 ± 0.08 | 0.77 ± 0.09 | 0.55 ± 0.14 | 0.63 ± 0.15 |
| | ($T_1$, $T_2$, $k_m$, ADC)* | 0.89 ± 0.04 | 0.85 ± 0.05 | 0.83 ± 0.11 | 0.77 ± 0.12 |
| Control vs. 16-h collagenase | ($T_1$, $T_2$) | 0.75 ± 0.08 | 0.73 ± 0.05 | 0.72 ± 0.13 | 0.66 ± 0.14 |
| | ($T_1$, $k_m$)* | 0.91 ± 0.03 | 0.92 ± 0.04 | 0.90 ± 0.08 | 0.90 ± 0.07 |
| | ($T_1$, ADC) | 0.73 ± 0.05 | 0.68 ± 0.06 | 0.70 ± 0.13 | 0.65 ± 0.13 |
| | ($T_2$, $k_m$) | 0.59 ± 0.11 | 0.55 ± 0.14 | 0.51 ± 0.16 | 0.45 ± 0.15 |
| | ($T_2$, ADC) | 0.57 ± 0.08 | 0.69 ± 0.09 | 0.51 ± 0.14 | 0.61 ± 0.16 |
| | ($k_m$, ADC) | 0.68 ± 0.08 | 0.48 ± 0.14 | 0.59 ± 0.15 | 0.38 ± 0.14 |
| | ($T_1$, $T_2$, $k_m$)* | 0.94 ± 0.04 | 0.96 ± 0.03 | 0.89 ± 0.08 | 0.92 ± 0.08 |
| | ($T_1$, $T_2$, ADC) | 0.79 ± 0.05 | 0.71 ± 0.06 | 0.75 ± 0.12 | 0.65 ± 0.12 |
| | ($T_1$, $k_m$, ADC)* | 0.95 ± 0.02 | 0.94 ± 0.03 | 0.89 ± 0.09 | 0.91 ± 0.08 |
| | ($T_2$, $k_m$, ADC) | 0.66 ± 0.07 | 0.63 ± 0.11 | 0.59 ± 0.13 | 0.48 ± 0.16 |
| | ($T_1$, $T_2$, $k_m$, ADC)* | 0.96 ± 0.04 | 0.96 ± 0.03 | 0.89 ± 0.08 | 0.90 ± 0.09 |

Analysis was performed using the same training and validation sets as for Table 2. Values shown are for the average of 100 independent selections of the training set and associated validation set.
* Indicates that classification using the indicated parameters results in improved sensitivity and specificity as compared to the optimal univariate analysis using $T_1$, for both the training and validation sets.

**Table 5**
Discriminant analysis of control and degraded BNC samples using multiple-component Gaussian clusters.

| Category | Parameters | Training set | | Validation set | |
|---|---|---|---|---|---|
| | | Sensitivity | Specificity | Sensitivity | Specificity |
| Control vs. 6-h trypsin | ($T_1$, $T_2$) | 0.75 ± 0.1 | 0.77 ± 0.1 | 0.65 ± 0.18 | 0.58 ± 0.18 |
| | ($T_1$, $k_m$)* | 0.88 ± 0.04 | 0.87 ± 0.06 | 0.79 ± 0.16 | 0.79 ± 0.15 |
| | ($T_1$, ADC) | 0.72 ± 0.1 | 0.78 ± 0.1 | 0.6 ± 0.17 | 0.6 ± 0.22 |
| | ($T_2$, $k_m$) | 0.74 ± 0.11 | 0.71 ± 0.12 | 0.51 ± 0.19 | 0.53 ± 0.19 |
| | ($T_2$, ADC) | 0.67 ± 0.11 | 0.78 ± 0.08 | 0.53 ± 0.17 | 0.66 ± 0.18 |
| | ($k_m$, ADC) | 0.75 ± 0.1 | 0.76 ± 0.09 | 0.58 ± 0.17 | 0.53 ± 0.18 |
| | ($T_1$, $T_2$, $k_m$)* | 0.94 ± 0.04 | 0.92 ± 0.06 | 0.78 ± 0.23 | 0.72 ± 0.26 |
| | ($T_1$, $T_2$, ADC) | 0.84 ± 0.12 | 0.89 ± 0.1 | 0.54 ± 0.32 | 0.64 ± 0.31 |
| | ($T_1$, $k_m$, ADC)* | 0.94 ± 0.04 | 0.95 ± 0.04 | 0.74 ± 0.29 | 0.72 ± 0.26 |
| | ($T_2$, $k_m$, ADC) | 0.81 ± 0.13 | 0.86 ± 0.11 | 0.52 ± 0.31 | 0.57 ± 0.27 |
| | ($T_1$, $T_2$, $k_m$, ADC) | 0.96 ± 0.05 | 0.96 ± 0.05 | 0.48 ± 0.39 | 0.79 ± 0.29 |
| Control vs. 16-h collagenase | ($T_1$, $T_2$) | 0.81 ± 0.05 | 0.79 ± 0.07 | 0.69 ± 0.17 | 0.62 ± 0.18 |
| | ($T_1$, $k_m$)* | 0.96 ± 0.03 | 0.95 ± 0.03 | 0.89 ± 0.1 | 0.86 ± 0.12 |
| | ($T_1$, ADC)* | 0.89 ± 0.05 | 0.87 ± 0.06 | 0.84 ± 0.12 | 0.77 ± 0.19 |
| | ($T_2$, $k_m$) | 0.67 ± 0.09 | 0.72 ± 0.11 | 0.5 ± 0.15 | 0.53 ± 0.17 |
| | ($T_2$, ADC) | 0.69 ± 0.1 | 0.84 ± 0.07 | 0.63 ± 0.16 | 0.71 ± 0.2 |
| | ($k_m$, ADC) | 0.72 ± 0.11 | 0.73 ± 0.15 | 0.6 ± 0.18 | 0.45 ± 0.18 |
| | ($T_1$, $T_2$, $k_m$)* | 0.98 ± 0.03 | 0.99 ± 0.02 | 0.79 ± 0.24 | 0.83 ± 0.23 |
| | ($T_1$, $T_2$, ADC) | 0.92 ± 0.06 | 0.92 ± 0.07 | 0.76 ± 0.27 | 0.68 ± 0.32 |
| | ($T_1$, $k_m$, ADC)* | 0.99 ± 0.02 | 0.99 ± 0.02 | 0.9 ± 0.19 | 0.79 ± 0.26 |
| | ($T_2$, $k_m$, ADC) | 0.81 ± 0.12 | 0.85 ± 0.09 | 0.63 ± 0.25 | 0.53 ± 0.26 |
| | ($T_1$, $T_2$, $k_m$, ADC)* | 0.99 ± 0.03 | 0.99 ± 0.03 | 0.74 ± 0.34 | 0.76 ± 0.32 |

Analysis was performed using the same training and validation sets as for Table 2. Values shown are for the average of 100 independent selections of the training set and associated validation set.
* Indicates that classification using the indicated parameters results in improved sensitivity and specificity as compared to the optimal univariate analysis using $T_1$, for both the training and validation sets.

The results were qualitatively and quantitatively comparable to those described above for discrimination between the set of controls and a set of samples degraded by either trypsin or collagenase. $T_1$ was again the best univariate classifier, while use of the parameter combinations ($T_1$, $k_m$), ($T_1$, $T_2$, $k_m$) and ($T_1$, $k_m$, ADC) resulted in markedly improved classification in both training and validation sets.

As discussed, fuzzy cluster analysis may be more applicable than binary classification to account for the variability in matrix composition across individual samples as well as the graded tissue degradation characteristic of osteoarthritis. Results from this approach are shown in Table 6. Since probabilistic assignments are given for each particular sample, aggregate sensitivity and specificity of assignment to a particular group is no longer the appropriate outcome measure. Table 6 therefore shows, for the control + trypsin-degraded samples, and separately for the control + collagenase-degraded samples, the probability of assignment of each validation set sample to the control or degraded group clusters, de-

**Table 6**
Cluster membership probabilities for validation samples.

| Parameters | Control + trypsin degraded | | | | Control + collagenase degraded | | | |
|---|---|---|---|---|---|---|---|---|
| | ($T_1$, $k_m$) | | ($T_1$, $k_m$, ADC) | | ($T_1$, $k_m$) | | ($T_1$, $k_m$, ADC) | |
| | Assigned to control group | Assigned to trypsin degraded group | Assigned to control group | Assigned to trypsin degraded group | Assigned to control group | Assigned to collagenase degraded group | Assigned to control group | Assigned to collagenase degraded group |
| Control samples | 0.78 | 0.22 | 0.85 | 0.15 | 0.99 | 0.01 | 1.00 | 0.00 |
| | 0.18 | 0.82 | 0.05 | 0.95 | 0.19 | 0.81 | 1.00 | 0.00 |
| | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 |
| | 0.98 | 0.02 | 0.99 | 0.01 | 1.00 | 0.00 | 1.00 | 0.00 |
| | 0.03 | 0.97 | 0.04 | 0.96 | 1.00 | 0.00 | 1.00 | 0.00 |
| | 0.83 | 0.17 | 0.80 | 0.20 | 1.00 | 0.00 | 0.99 | 0.01 |
| | 0.99 | 0.01 | 0.99 | 0.01 | 0.99 | 0.01 | 0.98 | 0.02 |
| | 0.99 | 0.01 | 0.99 | 0.01 | 1.00 | 0.00 | 1.00 | 0.00 |
| | 0.21 | 0.79 | 0.30 | 0.70 | 0.32 | 0.68 | 0.70 | 0.30 |
| | 0.33 | 0.67 | 0.34 | 0.66 | 0.58 | 0.42 | 0.60 | 0.40 |
| | 0.57 | 0.43 | 0.61 | 0.39 | 0.90 | 0.10 | 0.94 | 0.06 |
| | 0.98 | 0.02 | 0.99 | 0.01 | 0.97 | 0.03 | 0.99 | 0.01 |
| | 0.71 | 0.29 | 0.72 | 0.28 | 0.95 | 0.05 | 0.95 | 0.05 |
| | 0.40 | 0.60 | 0.46 | 0.54 | 0.95 | 0.05 | 0.98 | 0.02 |
| | 0.70 | 0.30 | 0.72 | 0.28 | 0.99 | 0.01 | 0.99 | 0.01 |
| Average | **0.64** | **0.36** | **0.66** | **0.34** | **0.85** | **0.15** | **0.94** | **0.06** |
| Degraded samples (Trypsin or Collagenase) | 0.17 | 0.83 | 0.27 | 0.73 | 0.11 | 0.89 | 0.39 | 0.61 |
| | 0.07 | 0.93 | 0.14 | 0.86 | 0.07 | 0.93 | 0.09 | 0.91 |
| | 0.16 | 0.84 | 0.28 | 0.72 | 0.06 | 0.94 | 0.02 | 0.98 |
| | 0.00 | 1.00 | 0.00 | 1.00 | 0.99 | 0.01 | 0.90 | 0.10 |
| | 0.07 | 0.93 | 0.09 | 0.91 | 0.04 | 0.96 | 0.01 | 0.99 |
| | 0.09 | 0.91 | 0.14 | 0.86 | 0.16 | 0.84 | 0.02 | 0.98 |
| | 0.30 | 0.70 | 0.41 | 0.59 | 0.16 | 0.84 | 0.09 | 0.91 |
| | 0.70 | 0.30 | 0.80 | 0.20 | 0.01 | 0.99 | 0.01 | 0.99 |
| | 0.04 | 0.96 | 0.06 | 0.94 | 0.10 | 0.90 | 0.00 | 1.00 |
| | 0.02 | 0.98 | 0.01 | 0.99 | 0.01 | 0.99 | 0.00 | 1.00 |
| | 0.50 | 0.50 | 0.33 | 0.67 | 0.04 | 0.96 | 0.00 | 1.00 |
| | 0.11 | 0.89 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 |
| Average | **0.19** | **0.81** | **0.21** | **0.79** | **0.15** | **0.85** | **0.30** | **0.70** |

Fuzzy clustering membership probabilities calculated as described in the text. The left four columns describe the results for the control + trypsin group, while the right four columns present findings for the control + collagenase group. For each of these, results are shown for each sample individually in a validation set consisting of 27 samples, with model parameters being defined from a training set of 53 samples. Results are shown for the pair ($T_1$, $k_m$) and ($T_1$, $k_m$, ADC), corresponding to sets providing favorable classification results (Tables 4 and 5). The averages of the calculated probabilities across samples are also presented, indicating the overall degree to which the samples were similar to controls or to degraded samples in terms of the parameter sets tested. Note that probability values for individual samples or averages which are not close to one or zero do not necessarily indicate classification difficulty or inadequacy of the model. Rather, such values are consistent with the graded nature of both the control and degraded samples.

fined according to the model-based discriminant approach. Out of the eleven possible parameter combinations, we chose the two that exhibited the greatest classification ability according to Table 5, namely ($T_1$, $k_m$) and ($T_1$, $k_m$, ADC). The centroids and model parameters of these clusters were again defined by 53 randomly-selected training set samples taken from the full group of controls + degraded samples. It is clear that the strength of the assignments varies among samples within a given group (control vs. trypsin-degraded; control vs. collagenase-degraded), reflecting tissue variability and the variability of the degradation, and the fact that MR parameter values vary continuously within both groups of samples. For example, the first control sample from the control + trypsin group was assigned with a probability of 0.78 to the control group, with a corresponding probability of 0.22 of assignment to the degraded group, according to the pair ($T_1$, $k_m$). The corresponding probabilities were 0.85 and 0.15 for classification according to the triplet ($T_1$, $k_m$, ADC). Similarly, the first degraded sample from the control + collagenase group was assigned with a probability of 0.11 to the control group, with a corresponding probability of 0.89 for assignment to the degraded group. The corresponding probabilities were 0.39 and 0.61 for assignment according to the triplet ($T_1$, $k_m$, ADC). The means of the assignment probabilities are also shown, and indicate the degree to which samples were assigned to groups according to the parameter sets indicated. We note that somewhat higher average assignment accuracy was achieved through use of the extended parameter set ($T_1$, $k_m$, ADC) as compared to ($T_1$, $k_m$) for the collagenase-degraded samples, indicating that the ADC measurement was sensitive to additional tissue manifestations of the degradative process. The addition of the ADC measurement was less effective for the trypsin degradation, indicating that the effect of this degradation on ADC was already adequately reflected in the ($T_1$, $k_m$) pair. Table 7 shows the summary results for all 11 parameter combinations; these again indicate the average probability of assignment to the indicated groups for the 27 validation samples. For example, using the parameter pair ($T_1$, $T_2$), the control samples were assigned to the control group with an average probability of 58% with the average assignment to the trypsin-degraded group therefore equaling 42%, while the trypsin-degraded samples were assigned to the trypsin-degraded group with an average probability of 59% with the average assignment to the control group therefore equaling 41%. It is important to emphasize that assignment probabilities substantially less than 100% do not necessarily reflect classification inaccuracies or other errors in the fuzzy clustering approach, but rather such values are to be expected and are consistent with actual sample heterogeneity.

## 4. Discussion

Noninvasive assessment of material properties is of substantial importance in biomedicine. Of available technologies, MRI is per-

**Table 7**
Means of cluster membership probabilities for validation samples.

| | | Assigned to control group | Assigned to trypsin degraded group | Assigned to control group | Assigned to collagenase degraded group |
|---|---|---|---|---|---|
| Control samples | ($T_1$, $T_2$) | 0.58 | 0.42 | 0.70 | 0.30 |
| | ($T_1$, $k_m$) | 0.64 | 0.36 | 0.85 | 0.15 |
| | ($T_1$, ADC) | 0.60 | 0.40 | 0.76 | 0.24 |
| | ($T_2$, $k_m$) | 0.52 | 0.48 | 0.54 | 0.46 |
| | ($T_2$, ADC) | 0.59 | 0.41 | 0.60 | 0.40 |
| | ($k_m$, ADC) | 0.53 | 0.47 | 0.57 | 0.43 |
| | ($T_1$, $T_2$, $k_m$) | 0.67 | 0.33 | 0.91 | 0.09 |
| | ($T_1$, $T_2$, ADC) | 0.63 | 0.37 | 0.77 | 0.23 |
| | ($T_1$, $k_m$, ADC) | 0.66 | 0.34 | 0.94 | 0.06 |
| | ($T_2$, $k_m$, ADC) | 0.60 | 0.40 | 0.64 | 0.36 |
| | ($T_1$, $T_2$, $k_m$, ADC) | 0.71 | 0.29 | 0.94 | 0.06 |
| Degraded samples (Trypsin or Collagenase) | ($T_1$, $T_2$) | 0.41 | 0.59 | 0.39 | 0.61 |
| | ($T_1$, $k_m$) | 0.19 | 0.81 | 0.15 | 0.85 |
| | ($T_1$, ADC) | 0.47 | 0.53 | 0.31 | 0.69 |
| | ($T_2$, $k_m$) | 0.54 | 0.46 | 0.54 | 0.46 |
| | ($T_2$, ADC) | 0.42 | 0.58 | 0.45 | 0.55 |
| | ($k_m$, ADC) | 0.49 | 0.51 | 0.55 | 0.45 |
| | ($T_1$, $T_2$, $k_m$) | 0.24 | 0.76 | 0.14 | 0.86 |
| | ($T_1$, $T_2$, ADC) | 0.40 | 0.60 | 0.28 | 0.72 |
| | ($T_1$, $k_m$, ADC) | 0.21 | 0.79 | 0.13 | 0.87 |
| | ($T_2$, $k_m$, ADC) | 0.48 | 0.52 | 0.51 | 0.49 |
| | ($T_1$, $T_2$, $k_m$, ADC) | 0.23 | 0.77 | 0.13 | 0.87 |

Average values for assignment probabilities of validation set samples for all eleven parameter combinations, calculated as in Table 6. Again, values between zero and one indicate sample variation, rather than inadequacy of the classification model.

haps the most widely applied due to the availability of a number of outcome measures which reflect a variety of biophysical characteristics. In this work we applied multiparametric cluster analysis to the problem of detecting early cartilage degradation. Such detection is an area of extensive ongoing investigation, and is regarded as a potentially essential component in the development of successful therapeutic approaches [22]. MRI-based studies of cartilage status often focus on changes in any one of several parameters, each of which is separately associated to a greater or lesser extent with particular aspects of cartilage degeneration. Characterization of cartilage using this approach in effect relies upon group mean changes in the parameter under investigation; that is, an observed parameter value is compared, explicitly or implicitly, to normative values for that parameter. However, this approach has met with limited success, with a great deal of overlap seen between parameter values observed for normal and diseased cartilage.

A study by Laurent et al. showed that while a 36-h collagenase digestion of both bovine nasal and articular cartilage resulted in 50% decreases in mean MTR and $k_m$, there was substantial overlap between the control and treatment groups, with the maximum values of MTR and $k_m$ in the collagenase-treated group greater than the smallest values in the control group for both types of tissue [4]. Similarly, comparable studies of the rabbit knee found that although mean $T_1$ increased with degradation, the maximum $T_1$ in control cartilage and the minimum $T_1$ of papain-degraded cartilage were ~680 and ~480 ms, respectively [7]. Indeed, half of the total data points obtained were within this region of $T_1$ overlap. In a study of chondroitinase-ABC treatment of neocartilage developing under a tissue engineering protocol, Chen et al. found that although the maximum transverse relaxation rate $R_2 = 1/T_2$ of controls was 25% higher than that of chondroitinase treated samples, the minimum $R_2$ values were the same for the two groups [8]. It was also found that approximately half of the fixed charged density (FCD) values of the chondroitinase-treated samples were close to the FCD mean of the control group.

In view of the limitations of univariate classification, we undertook a study of multiparametric classification of cartilage under pathomimetic degradation protocols. Multiparametric analyses allow several variables to be considered simultaneously; in addition, they provide quantitative relationships between different measured parameters, as well as quantitative measures of data cluster morphology and cluster movement secondary to intervention. Related methodologies have been widely applied to MR analysis of other tissues such as breast and brain [23,24]. We studied several approaches, including classification based on the arithmetic means of MRI parameters, multiparametric *k*-means clustering, and model-based discriminant analyses, restricted to single-component clusters and without such restriction. Sensitivity and specificity to degradation were evaluated quantitatively for each of these for trypsin- and collagenase-digested samples.

Our main interest is in the effectiveness of the multiparametric cluster approach to detect early stages of cartilage degradation, when other analyses such as morphologic evaluation and unvariate analysis are of limited value. We therefore used relatively mild enzymatic degradation protocols. This contrasts with our previous study [14], in which a greater degree of cartilage degradation was induced through use of longer incubation times, as evidenced by larger changes in MR parameters from controls in the degraded tissues (compare Table 1 of Ref. [14] with Table 1 of the present study). In the setting of more pronounced degradation, certain univariate analyses achieved fairly high degrees of accuracy [14]. However, with milder degradation as employed in the present manuscript, the accuracy of the univariate approach was much more limited, as shown by comparing Table 2 of Ref. [14] with Table 2 of the present study.

Although BNC exhibits a greater degree of homogeneity than does articular cartilage, tissue inhomogeneity was evident both prior to, and especially after, enzymatic degradation; this is evident in Fig. 1b. The general magnitude of these effects can be determined from previous Fourier transform imaging spectroscopic studies of BNC [25]. In that study, trypsin degradation resulted in CV of 24% for the spatial distribution of collagen, and 43% for PG distribution. The inhomogeneities resulting from biological variation in the control tissue and nonuniformity of sample degradation, which contribute to an overlap in the data between the control and degraded groups, present an additional challenge to the classification analysis. However, such heterogeneities are also a given even in non-degraded articular cartilage due to its layered architecture,

with further heterogeneity being characteristic of osteoarthritic cartilage. Thus, the success of the classification scheme implemented here indicates its likely applicability to articular cartilage as well.

Our approach of multiple sample imaging led to use of a large imaging volume, so that $B_1$ inhomogeneity is also evident. We did not expect this image shading to affect the derivation of MR parameters based on fitted data curves, and indeed, analysis indicated that parameters derived from samples exhibiting shading did not differ in any systematic way from those derived from samples located more centrally in the coil.

### 4.1. Classification based on arithmetic means and multivariate k-means clustering

Although changes in mean MR parameter values were observed upon degradation with both trypsin and collagenase, there was a large degree of overlap between group values, resulting in limited classification accuracy using the arithmetic mean of any one parameter. We then implemented *k*-means clustering as one of the simplest and most widely applied approaches to multiparametric classification. However, even with the inclusion of up to four MRI parameters, *k*-means clustering did not yield greater classification accuracy than did simple arithmetic means. We again attribute this to parameter overlap, as well as the inability of this algorithm to distinguish between different cluster morphologies. An additional limitation of the *k*-means algorithm is its sole reliance on minimization of the Euclidian distance between sample point positions and cluster centroids, resulting in convergence to uniform density hyperspherical clusters. This imposes a strong constraint on the types of datasets most appropriately modeled, with poorer performance therefore expected for non-uniform density, non-hyperspherical, data clusters, such as seen in the present case.

### 4.2. Classification based on model-based clustering

In contrast to the *k*-means approach, the model-based clustering algorithm implemented here assigns clusters based on maximizing the probability of obtaining the observed data as a function of cluster assignment [17]. The models available through MCLUST permit independent variation of cluster component volumes, shapes, and orientations; this approach also permits the inclusion of more than one independent Gaussian component for a given cluster, as illustrated in Fig. 3. For the particular data represented in that figure, the full model-based procedure identified two ellipsoidal components for both the control samples and the trypsin-degraded samples.

The ability to classify the data through selection of multiple parameters characterizing the clusters resulted in improved sensitivity and specificity. Because the models are developed to most closely fit the training set, there is generally more misclassification in the validation set than in the training set. For the same reason, model parameters resulting in the best classification rates for the training set do not always result in the best possible classification accuracy in the validation set.

We performed additional analyses in which the ratio of the size of the training set to the size of the validation set was varied. We found that results for validation set classification improved as the proportional size of the training set increased from 1/4 to approximately 2/3 of the total number of samples. Results did not improve for training sets comprised of proportions of the data greater than this. We note that results for the validation set as defined by our analysis may underestimate actual sensitivity and specificity, since the training set represents only a subset of the full data set [16]. Nevertheless, error rates for the ($T_1$, $k_m$), ($T_1$, $T_2$, $k_m$) and ($T_1$, $k_m$, ADC) parameter combinations demonstrated classification accuracy beyond the results for any of the univariate analyses.

The classification approaches described here indicate the potential for multiparametric detection of early cartilage degeneration. Other approaches may be of additional benefit and remain to be explored. The assumption of Gaussian sample point distribution can be relaxed through use of non-Gaussian clustering algorithms [26]. The potential advantages of such an approach, especially when weighed against the additional complexity of the analysis, would depend upon more formal analysis of sample point distribution in parameter space. Further, we note that clinical diagnostic studies may depend upon establishing multiparametric data sets for normative tissue.

We used ROI's which were voxel averages for MRI parameter determination, in effect increasing the SNR of the data to be fit. Indeed, data averaging is one of a number of experimental details determining the SNR of a data set, including, for example, voxel size (including the effects of voxel averaging), field strength, coil quality, and number of acquisitions. A limitation of our study is that we did not perform a formal analysis of the SNR requirements for given degrees of classification accuracy. However, no significant difference in SNR was observed between control and degraded samples. Similarly, although the BNC samples studied in the present manuscript do not exhibit the anisotropy characteristic of articular cartilage, we are not aware of magnetic resonance studies properly documenting changes in cartilage anisotropy as a function of degradation. Therefore, it is not clear to what extent the more organized structure of articular samples would impact, or provide an additional parameter of interest for, the multiparametric approach.

Although the emphasis in the present study was on the exploration of multiparametric analytic techniques, we note that our univariate analysis indicated that of single parameters, classification according to $T_1$ showed the greatest accuracy. This somewhat surprising result was also found in an extensive comparable analysis of more strongly degraded cartilage [14]. This finding appears to be attributable to the relatively small CV of the observed $T_1$ values. Variability in results for all parameters is expected to result from intrinsic sample inhomogeneity, spatial and temporal variability of enzymatic degradation, and finite precision of the MRI experiment itself. The relatively low CV of the $T_1$ measurement, then, may be due to the relative lack of sensitivity of $T_1$ to small cartilage matrix variations.

In order to extend the discriminant analysis to graded tissue properties, such as is characteristic of degenerative cartilage in general, we implemented a fuzzy-clustering procedure. Classification models were again established using a training set, but each sample in the validation sets was individually assigned in a probabilistic fashion to either the control or degraded group. As expected with samples exhibiting both biological variability and varying degrees of degradation due, for example, to differing sample sizes, the sample-by-sample assignments vary across a range of probabilities. We calculated mean assignment probabilities across the 27 validation samples for each of the 11 parameter sets; the results show consistency but not complete agreement. This indicates that the probability of assignment to a control or a degraded group is a function of the parameters evaluated, as would be expected. We note that a fuzzy cluster approach to tissue classification would also be appropriate for the analysis of articular cartilage, which is heterogeneous even absent degradation.

In conclusion, the sensitivity and specificity of a hierarchy of data clustering techniques to pathomimetic degradation of cartilage were evaluated. Univariate classification, as is implicitly used in most contemporary studies, as well as *k*-means clustering exhibited limited ability to discriminate between control and degraded tissue. Use of multiparametric model-based discriminant analysis, on the other hand, resulted in substantial improvement in classifi-

cation accuracy, due to the ability of those models to more accurately reflect sample distribution in parameter space. The ability to classify samples with substantially improved accuracy has direct implications for improved detection of regions of degradation in MR images, so that multiparametric analysis may lead to substantially improved ability to detect early cartilage degradation. Finally, we note that our results support the applicability of this approach to the nondestructive evaluation of other biomaterials using MRI.

## Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Institute on Aging.

## References

[1] G.E. Gold, C.F. Beaulieu, Future of MR imaging of articular cartilage, Semin. Musculoskelet. Radiol. 5 (2001) 313–327.
[2] T.J. Mosher, B.J. Dardzinski, Cartilage MRI T2 relaxation time mapping: overview and applications, Semin. Musculoskelet. Radiol. 8 (2004) 355–368.
[3] D. Burstein, A. Bashir, M.L. Gray, MRI techniques in early stages of cartilage disease, Invest. Radiol. 35 (2000) 622–638.
[4] D. Laurent, J. Wasvary, J. Yin, M. Rudin, T.C. Pellas, E. O'Byrne, Quantitative and qualitative assessment of articular cartilage in the goat knee with magnetization transfer imaging, Magn. Reson. Imaging 19 (2001) 1279.
[5] N.M. Menezes, M.L. Gray, J.R. Hartke, D. Burstein, T2 and T1ρ MRI in articular cartilage systems, Magn. Reson. Med. 51 (2004) 503–509.
[6] V. Mlynarik, I. Sulzbacher, M. Bittsansky, R. Fuiko, S. Trattnig, Investigation of apparent diffusion constant as an indicator of early degenerative disease in articular cartilage, J. Magn. Reson. Imaging 17 (2003) 440–444.
[7] D. Laurent, J. Wasvary, E. O'Byrne, M. Rudin, In vivo qualitative assessments of articular cartilage in the rabbit knee with high-resolution MRI at 3 T, Magn. Reson. Med. 50 (2003) 541–549.
[8] C.-T. Chen, K.W. Fishbein, P.A. Torzilli, A. Hilger, R.G.S. Spencer, W.E. Horton Jr., Matrix fixed-charge density as determined by magnetic resonance microscopy of bioreactor-derived hyaline cartilage correlates with biochemical and biomechanical properties, Arthritis Rheum. 48 (2003) 1047–1056.
[9] M.T. Nieminen, J. Toyras, J. Rieppo, J.M. Hakumaki, J. Silvennoinen, H.J. Helminen, J.S. Jurvelin, Quantitative MR microscopy of enzymatically degraded articular cartilage, Magn. Reson. Med. 43 (2000) 676–681.
[10] R. Meder, S.K. de Visser, J.C. Bowden, T. Bostrom, J.M. Pope, Diffusion tensor imaging of articular cartilage as a measure of tissue microstructure, Osteoarthritis Cartilage 14 (2006) 875–881.
[11] A. Berg, T. Singer, E. Moser, High-resolution diffusivity imaging at 3.0 T for the detection of degenerative changes – a trypsin-based arthritis model, Invest. Radiol. 38 (2003) 460–466.
[12] C. Fraley, A.E. Raftery, MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering. Technical Report 504, Department of Statistics, University of Washington, 2006.
[13] J.V. Hajnal, C.J. Baudouin, A. Oatridge, I.R. Young, G.M. Bydder, Design and implementation of magnetization transfer pulse sequences for clinical use, J. Comput. Assist. Tomogr. 16 (1992) 7–18.
[14] P.-C. Lin, D.A. Reiter, R.G. Spencer, Sensitivity and specificity of univariate MRI analysis of experimentally degraded cartilage, Magn. Reson. Med., 2009, doi:10.1002/mrm.22110.
[15] D.A. Reiter, P.C. Lin, K.W. Fishbein, R.G. Spencer, Multicomponent T2 relaxation analysis in cartilage, Magn. Reson. Med. 61 (2009) 803–809.
[16] A.M. Molinaro, R. Simon, R.M. Pfeiffer, Prediction error estimation: a comparison of resampling methods, Bioinformatics 21 (2005) 3301–3307.
[17] C. Fraley, A.E. Raftery, Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST, J. Classif. 20 (2003) 263–286.
[18] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a k-means clustering algorithm, Appl. Stat. 28 (1979) 100–108.
[19] R: A language and environment for statistical computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2005.
[20] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. R. Stat. Soc. B Methodol. 58 (1996) 155–176.
[21] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[22] M.P. Recht, D.W. Goodwin, C.S. Winalski, L.M. White, MRI of articular cartilage: revisiting current status and future directions, Am. J. Roentgenol. 185 (2005) 899–914.
[23] G. Harris, N.C. Andreasen, T. Cizadlo, J.M. Bailey, H.J. Bockholt, V.A. Magnotta, S. Arndt, Improving tissue classification in MRI: a three-dimensional multispectral discriminant analysis method with automated training class selection, J. Comput. Assist. Tomogr. 23 (1999) 144–154.
[24] F. Forbes, N. Peyrard, C. Fraley, D. Georgian-Smith, D.M. Goldhaber, A.E. Raftery, Model-based region-of-interest selection in dynamic breast MRI, J. Comput. Assist. Tomogr. 30 (2006) 675–687.
[25] K. Potter, L.H. Kidder, I.W. Levin, E.N. Lewis, R.G.S. Spencer, Imaging of collagen and proteoglycan in cartilage sections using Fourier transform infrared spectral imaging, Arthritis Rheum. 44 (2001) 846–855.
[26] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Stat. Assoc. 97 (2002) 611–631.