Magnetic Resonance Imaging of Inflammatory Lesions in the Spine in Ankylosing Spondylitis Clinical Trials: Is Paramagnetic Contrast Medium Necessary?

KAY-GEERT A. HERMANN, ROBERT B.M. LANDEWÉ, JÜRGEN BRAUN, and DÉSIRÉE M.F.M. van der HEIJDE

ABSTRACT. Depiction of inflammatory lesions by magnetic resonance imaging (MRI) in ankylosing spondylitis (AS) is possible both by short-tau inversion recovery (STIR) imaging and by gadolinium-enhanced T1-weighted imaging with fat saturation (T1/Gd). The aim of this prospective study was to investigate whether Gd-enhanced sequences add relevant information compared to STIR imaging alone in the detection of active spinal lesions. MRI of the spine was performed in 48 patients with AS, who participated in a clinical trial of tumor necrosis factor blocking drugs, by STIR and T1/Gd at baseline and after 6 months. Images were evaluated separately for the 2 techniques by 2 readers blinded for true time sequence and treatment. The ASspiMRI-a scoring method was used, in which 23 vertebral units are graded for inflammation from 0 to 6 (total score 0 to 138). Mean scorings of both techniques within readers were in the same range (reader 1: STIR 7.8, T1/Gd 7.7; reader 2: STIR 4.4, T1/Gd 4.7). Intraclass correlation coefficients comparing STIR and T1/Gd where high for both status scores (reader 1: 0.88; reader 2: 0.90) and change scores (both readers: 0.88). Bland and Altman analysis for both sequences showed homogeneous interreader variability along the entire spectrum of scorings, for both status scores and change scores. Smallest detectable change for status scores was 6.2 for STIR and 6.7 for T1/Gd, and for change scores 6.5 and 6.3, respectively. Standardized response means were comparable for both methods (range: 0.80–1.09). In conclusion, both STIR and T1/Gd sequences measure inflammation of the spine, as well as change of inflammation, with a high level of agreement between the 2 sequences. For future clinical randomized trials with MRI of the spine as outcome measure, STIR could be considered for use as the sole imaging technique. (J Rheumatol 2005;32:2056–60)

Key Indexing Terms:
ANKYLOSING SPONDYLITIS MAGNETIC RESONANCE IMAGING SPINE
INFLAMMATION STIR Gd-DTPA ASspiMRI-a

Introduction

Ankylosing spondylitis (AS) is a chronic inflammatory rheumatic disease that mainly affects the axial skeleton. The gold standard of assessing structural damage in AS is radiography of spine and pelvis, with the ability to detect chronic changes like syndesmophytes[1]. However, there is increasing evidence that magnetic resonance (MR) imaging is able to detect acute spinal lesions even in the early stages of the disease, and to assess changes in such lesions over time in patients treated with tumor necrosis factor-α (TNF-α) targeting therapy[2,3].

F*rom the Department of Radiology, Charité Medical School, Berlin, Germany; Department of Internal Medicine/Rheumatology, University Hospital Maastricht, Maastricht, The Netherlands; and Rheumazentrum Ruhrgebiet, Herne, Germany.*

K-G.A. Hermann, MD, Department of Radiology, Charité Medical School; R.B.M. Landewé, MD, PhD, Department of Internal Medicine/ Rheumatology, University Hospital Maastricht; J. Braun, MD, Rheumazentrum Ruhrgebiet; D.M.F.M. van der Heijde, MD, PhD, Department of Internal Medicine/Rheumatology, University Hospital Maastricht.

Address reprint requests to Dr. K-G.A. Hermann, Department of Radiology, Charité University Hospital, Campus Mitte, Schumannstrasse 20/21, 10117 Berlin, Germany. E-mail: kgh@charite.de

Using MR imaging, T1-weighted sequences are usually applied to evaluate chronic changes. For the detection of acute changes a variety of techniques can be used: Sequence 1, short-tau inversion recovery (STIR) sequence; Sequence 2, T2-weighted fat suppressed fast spin echo (FSE) sequence (T2/FS); and Sequence 3, T1-weighted fat suppressed FSE sequence after administration of paramagnetic contrast medium such as gadolinium diethylenetriamine pentaacetic acid (Gd-DTPA) (T1/Gd). In this article sequences 1 and 3 are analyzed. Acute spinal lesions are depicted as bright spots, whereas normal bone marrow appears dark in all 3 sequences[4]. There are, however, different underlying technical principles. Both T2/FS and STIR sequences rely on the high T2 relaxation time of free water, e.g., in bone marrow edema, and represent it as bright areas. In contrast, T1/Gd sequences depict inflammatory lesions that are bright due to their higher vascularity on the basis of diffusion of Gd-DTPA molecules into the interstitium. The major disadvantage of the T1/Gd technique is the need for injection of contrast medium, which makes MR imaging invasive, more time consuming, and more expensive. The STIR sequence, on the other hand, has a lower signal to

noise ratio. In the setting of a clinical trial, imaging protocols should be very straightforward and easy to perform in different centers.

We evaluated whether in clinical trials T1-weighted fat saturated contrast enhanced sequences add relevant information compared to STIR imaging alone for the depiction of acute spinal lesions in AS.

MATERIALS AND METHODS

Patients. Forty-eight patients with a diagnosis of AS according to the modified New York classification criteria[5] were randomly selected from 279 patients who participated in a double-blind placebo controlled trial comparing infliximab as TNF-α targeting agent or placebo[6]. The median age of the total trial population was 40 years, the median disease duration was about 12 years. The kind of medication was blinded to the readers.

MR imaging protocol. MR imaging was performed at baseline and after 6 months. No specific MR device was applied; rather, a variety of machines were used since image acquisition was performed at several centers. All imaging centers used whole-body magnets with a field strength of 1.0 Tesla or 1.5 Tesla. Patients were in a supine position and a spine array coil was used.

The whole spine was imaged starting with the upper part, including the cervical spine and the upper portion of the thoracic spine, and afterwards the lower spine including the lower part of the thoracic spine and the lumbar spine. Special care was taken to ensure sufficient overlap between imaging of upper and lower parts of the spine. The following sequences were used as described[2], all in sagittal section orientation: (1) T1-weighted fast spin echo (FSE) sequence (repetition time, TR: 500 ms, echo time, TE: 44 ms); (2) STIR sequence (TR: 4420 ms, TE: 79 ms, inversion time: 150 ms); and (3) T1-weighted FSE (TR: 599 ms, TE: 44 ms) with fat saturation after administration of Gd-containing contrast medium with a dosage of 0.1 mmol/kg body weight. The field of view was 380 mm, slice thickness 4 mm, and matrix size 512 × 256 pixels in all sequences. Whereas sequences 2 and 3 where analyzed in detail for the purpose of this study, sequence 1 served only for definition of morphology in doubtful cases, especially to determine the presence of erosions.

Scoring. The scoring of the images was performed by one rheumatologist (RL) and one radiologist (KGH) experienced in musculoskeletal MR imaging. As our scoring method we used the ASspiMRI-a, an AS spinal MRI scoring system that was recently proposed[2]. This method was found to be reliable with low intra-rater variance (5.3–7.7)[2]. Briefly, acute changes are scored on the basis of the area of bone marrow edema per vertebral unit (VU). A VU is defined as the region between 2 virtual lines drawn parallel to the vertebral endplates through the middle of each vertebra. Scores from 1 to 3 comprise acute changes that present only bone marrow edema (1: up to 25% of area; 2: up to 50% of area; 3: > 50% of area of VU); scores from 4 to 6 are based on spinal lesions showing both edema and erosions (4: minor erosion with bone marrow edema; 5: moderate erosion with bone marrow edema; 6: severe erosion with bone marrow edema). Erosions without bone marrow edema are considered inactive, chronic changes that must not be included in the ASspiMRI-a score. All 23 VU between C2 and S1 are included in the ASspiMRI-a score, which sums to a total score of 138. Involvement of a VU was assumed if the score was ≥ 1. Both readers were trained in using the scoring system.

In addition, a quality score was assigned to each sequence with the following definition: 1: very poor image quality, unable to read; 2: severely impaired quality, but able to read; 3: moderate image quality; 4: good image quality with few artifacts; and 5: excellent image quality.

Readings were done independently by both readers blinded to the time sequence and to type of treatment. Immediately afterwards, they scored the T1/Gd sequence, without comparing it directly to the STIR sequence. No adjustments of STIR reading scores were allowed after seeing the T1/Gd sequence. Technical limitations of the reading systems did not allow for a different setting.

Statistical analysis. The data of both readers were evaluated separately since comparison of STIR versus T1/Gd was the main interest of the study. Statistical analysis first comprised descriptive measures such as mean sum scores and mean change scores (change defined as absolute value of the difference between baseline and followup score) and number of affected VU. Overall correlation of the 2 sequences was calculated by Spearman rank correlation.

In order to investigate agreement of scores obtained by STIR and T1/Gd, intraclass correlation coefficients (ICC; absolute agreement definition) were calculated per reader, for both status scores and change scores.

In order to visualize patterns of agreement between both readers per sequence, Bland and Altman plots[7,8] were constructed, for both status scores and change scores, and smallest detectable changes (SDC) were indicated. SDC is defined as 1.96 × (standard deviation (SD) of the interreader differences)/$\sqrt{2}$[9]. Interreader agreement per sequence was compared by SDC, favoring the sequence with the lower SDC value.

In order to compare sensitivity to change of STIR versus T1/Gd, standardized response means (SRM) were calculated on the absolute scores (without taking the direction of the change into account). The latter was done because the real time order of the MR images was unknown. The SRM is defined as the mean change in score divided by the standard deviation of the change in scores[10,11].

RESULTS

Descriptive results. At least one active inflammatory lesion was present in 43 of the 48 patients for reader 1 and in 34 of the 48 patients for reader 2 in the STIR or T1/Gd sequence. On average, reader 1 scored higher than reader 2 in both sequences. Scorings of both MR imaging techniques within readers were in the same range (see Table 1). There was no clear preference for either technique, since reader 1 scored STIR images slightly higher, whereas reader 2 scored T1/Gd higher. The same was true for the general impression of the quality of images. The median quality score was similar for both sequences, for both readers. The median number of affected VU was about 4 for reader 1 and about 1 for reader 2 for both sequences (Table 1). Erosions, i.e., scores ≥ 4, in the ASspiMRI-a scoring system, were present in only a few patients, with a very low number of VU with erosions per patient.

Method comparison. For both status scores and change scores intraclass correlation coefficients comparing STIR and T1/Gd were high for both readers (Table 1). On a patient level, the rate of concordance was 89.3% for reader 1 and 87.2% for reader 2. For reader 1, this means that 2.1% of patients would have been scored normal by T1/Gd but abnormal by STIR score, and 8.5% vice versa. For reader 2 these values are 8.5% and 4.3%, respectively. Spearman's rho correlation coefficients between STIR and T1/Gd were 0.87 for reader 1 and 0.83 for reader 2. On the level of vertebral units, i.e., taking all VU together, there where high rates of concordance too: 87.6% for reader 1, with 6.2% normal using T1/Gd and 6.3% normal using STIR; and 94.9% for reader 2, with 1.7% normal in T1/Gd and 3.4% not affected in STIR.

Bland and Altman analysis showed for both sequences

Table 1. Summary of all results for the 2 readers. Values in parentheses are medians with 25th and 75th quartile, unless stated otherwise.

	Reader 1		Reader 2	
	STIR	T1/Gd	STIR	T1/Gd
Mean score (SD)	7.8 (7.0)	7.7 (6.6)	4.4 (5.9)	4.7 (6.2)
No. of affected VU per patient	4 (1; 6.8)	3.5 (1.3; 5)	1 (0; 3.8)	1.5 (0; 4)
No. of affected VU with erosions per patient	0 (0; 0.8)	0 (0; 0)	0 (0; 0)	0 (0; 0)
ICC of status scores (95% CI)	0.88 (0.79–0.93)		0.90 (0.83–0.94)	
ICC of change scores (95% CI)	0.88 (0.80–0.93)		0.88 (0.79–0.93)	
SRM	1.09	0.97	0.80	0.88
Quality score	4 (2; 5)	4 (3; 5)	4 (3; 5)	4 (4; 5)

STIR: short-tau inversion recovery sequence; T1/Gd: T1-weighted sequence fat suppressed after gadolinium administration; VU: vertebral units; ICC: intraclass correlation coefficient; SRM: standardized response mean.

that interreader variability was homogeneous along the entire spectrum of scorings, both for scoring status and for scoring change, with a consistent (the same direction in both sequences) systematic difference (Figure 1, broken line) with respect to status scores, and a less prominent systematic difference with respect to change scores. Patterns of both sequences were entirely comparable. The SDC for status score, based on the agreement between readers, was 6.7 for STIR and 6.2 for T1/Gd (Figure 1A and 1B). The SDC for change score was 6.3 for STIR and 6.5 for T1/Gd (Figure 2A and 2B). Thus very few differences exist between both MR sequences with regard to inter-reader variability and SDC.

Sensitivity to change. SRM were roughly similar for both methods, but reader 1 consistently reached higher SRM than reader 2 (Table 1).

DISCUSSION

We aimed to perform a direct comparison of STIR and T1/Gd sequences for use in AS clinical trials. This matter becomes more important since MR imaging is becoming the new gold standard for detection of acute changes in AS. By omitting the application of contrast material, MR examinations would be less costly and — more important perhaps — more feasible.

This report is one of the first to compare STIR and T1/Gd sequences in inflammatory rheumatic disease of the spine. We used the ASspiMRI-a scoring system, which was developed by our group and has proven to be reliable[2], to compare the 2 imaging techniques. Comparing several validity aspects of the scorings in both readers separately revealed that both sequences behaved similarly. ICC comparing status and change scores were excellent. Somewhat in contrast

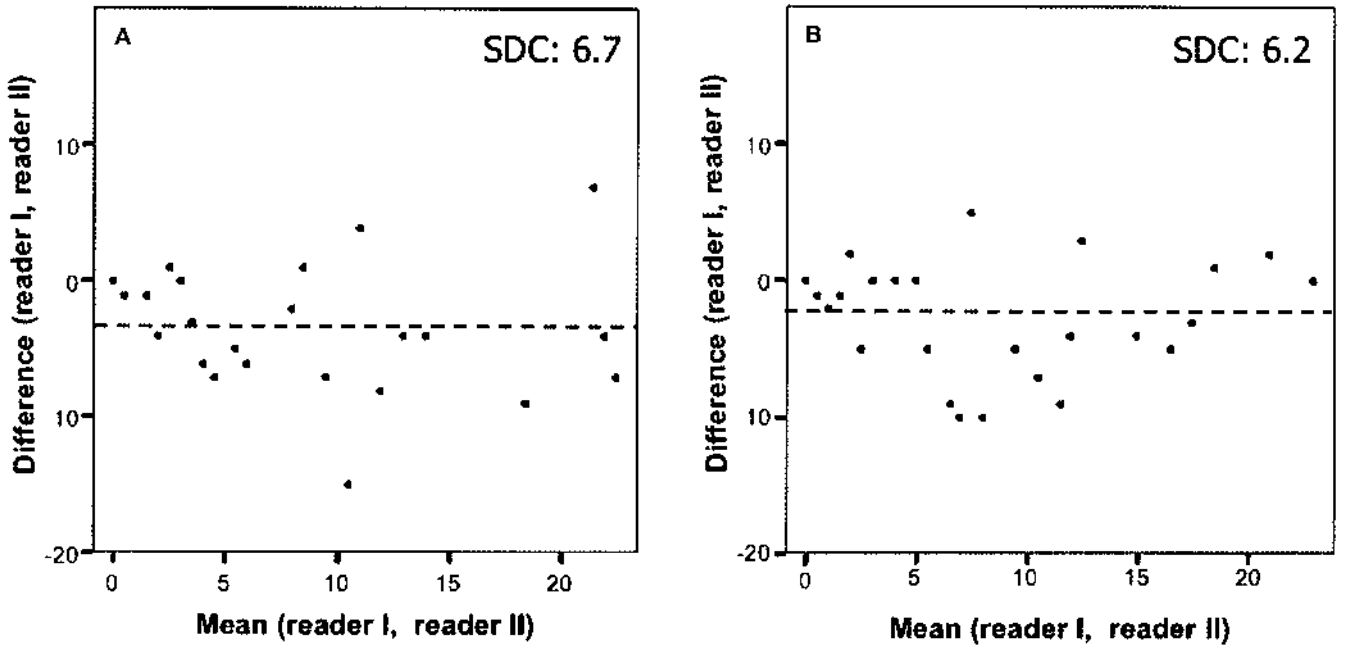

Figure 1. Bland and Altman plots. A. Distribution of status scores for short-tau inversion recovery (STIR) sequence. B. Distribution of status scores for gadolinium-enhanced T1-weighted sequence with fat saturation (T1/Gd). SDC: smallest detectable change.

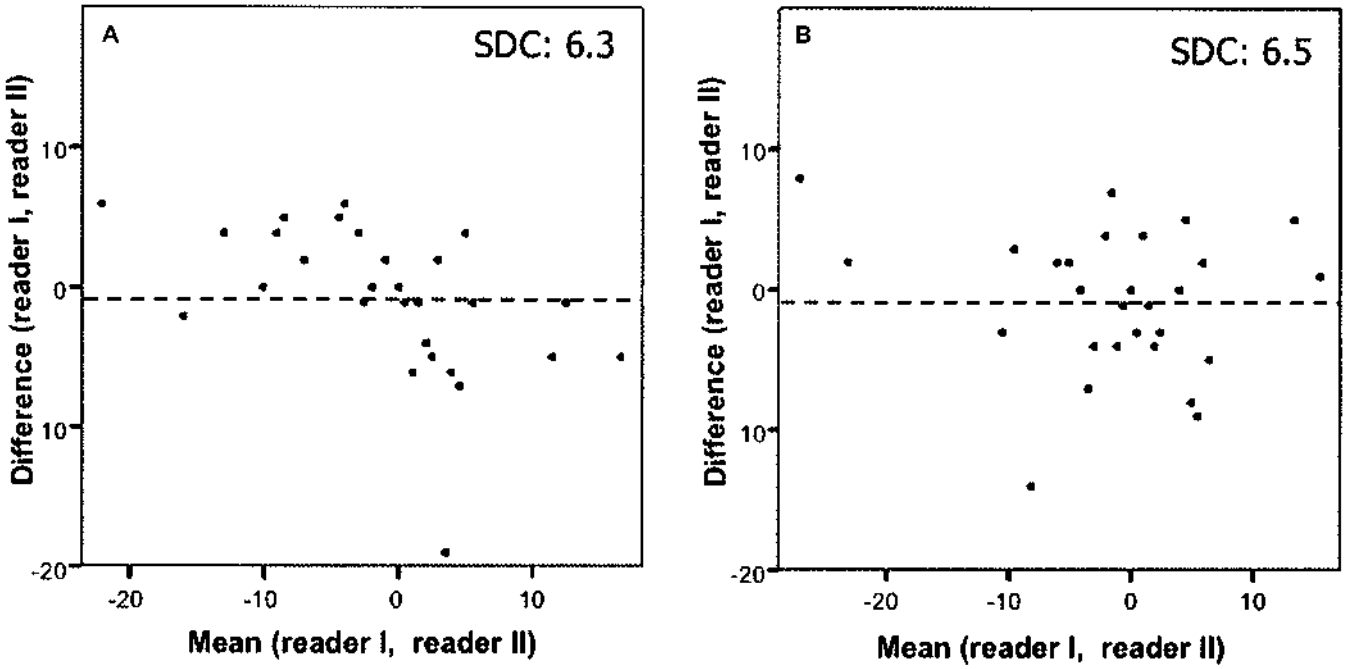

Figure 2. Bland and Altman plots. A. Distribution of change scores for STIR sequence. B. Distribution of change scores for T1/Gd sequence. SDC: smallest detectable change.

to what was expected, T1/Gd sequences do not seem to add significant information, since the number of affected VU found by T1/Gd was only slightly higher than for STIR. Similar results were obtained in our previous study based on a smaller number of patients[2]. In that study, we also presented the results separately for STIR and T1/Gd, but we arrived at a Spearman correlation coefficient for performance of both sequences of only 0.59. An ICC, which would have allowed better comparison with our current study, was not calculated, and would have provided a somewhat lower value. In the current study, we achieved correlation coefficients of up to 0.87. Explanations may be the extensive training aiming at consensus scoring of abnormalities, as well as differences in image quality. The high quality scores obtained in this study add to the latter explanation. High correlations similar to those in the present study were found by our group in a different study[12].

The value of STIR images was outlined by a number of authors previously, e.g., in multiple myeloma[13,14] and vertebral fractures[15]. All these studies found STIR images to be the most sensitive for the pathology in question, and T1/Gd images either did not perform better or did not add relevant information.

In our present study, we did not include fat saturated T2-weighted images for comparison with STIR and T1/Gd images. Some reports have addressed this issue, mainly in malignant bone marrow disease of the spine. Jones, *et al* found no significant differences either in contrast-to-noise ratios or in detection rates of metastatic spinal lesions[16]. Focal lesions in multiple myeloma were detected equally well by both STIR and T2-weighted sequences by Rahmouni, *et al*[13], whereas Baur, *et al* found STIR sequences to be more sensitive[14].

Although carefully planned, our study has some limitations. First, the type of treatment (infliximab versus placebo) was not disclosed. Knowing this would have allowed for proper calculation of SRM separately for patients treated with infliximab or placebo. Second, the time sequence of MR images was unknown, and therefore only absolute values of change could be taken into account. Thus, improvement or deterioration of scoring results could not be analyzed for the different MR sequences. A third limitation is the technical setup used. The workstation for reading the images allowed only one opportunity to call up data for a patient. Consequently, both sequences were read in a close timeframe. However, only one type of sequence was displayed at a time, and scoring results were not changed by readers after entry in the scoring sheet.

It should be emphasized that conclusions of this study refer to use of MRI in the context of clinical trials, following and comparing groups of patients, not to individual patients and/or diagnostic problems. We did not compare discrimination between presence and absence of disease, since all patients had AS. We did not investigate sensitivity of STIR and T1/Gd with regard to activity in single VU, since sum scores were compared, and we did not investigate specificity of abnormalities detected with STIR and/or T1/Gd. So undoubtedly T1/Gd may be of added value in situations other than clinical trials.

In conclusion, for future clinical randomized trials with

magnetic resonance imaging of the spine as outcome parameter it could be considered to use only STIR as an imaging technique for the detection of change in acute lesions.

REFERENCES

1. Spoorenberg A, de Vlam K, van der Linden S, et al. Radiological scoring methods in ankylosing spondylitis. Reliability and change over 1 and 2 years. J Rheumatol 2004;31:125-32.
2. Braun J, Baraliakos X, Golder W, et al. Magnetic resonance imaging examinations of the spine in patients with ankylosing spondylitis, before and after successful therapy with infliximab: evaluation of a new scoring system. Arthritis Rheum 2003;48:1126-36.
3. Marzo-Ortega H, McGonagle D, O'Connor P, Emery P. Efficacy of etanercept in the treatment of the entheseal pathology in resistant spondylarthropathy: a clinical and magnetic resonance imaging study. Arthritis Rheum 2001;44:2112-7.
4. Hermann KGA, Bollow M. Magnetic resonance imaging of the axial skeleton in rheumatoid disease. Best Pract Res Clin Rheumatol 2004;18:881-907.
5. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.
6. van der Heijde D, Dijkmans B, Geusens P, et al. Ankylosing spondylitis study for the evaluation of recombinant infliximab therapy study group. Efficacy and safety of infliximab in patients with ankylosing spondylitis: results of a randomized, placebo-controlled trial (ASSERT). Arthritis Rheum 2005;52:582-91.
7. Bland JM, Altman DG. Measurement error [statistics notes]. BMJ 1996;313:744.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
9. Bruynesteyn K, Boers M, Kostense P, van der Linden S, van der Heijde D. Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change? Ann Rheum Dis 2005;64:179-82. Epub 2004 Jul 29.
10. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632-42.
11. Angst F, Aeschlimann A, Steiner W, Stucki G. Responsiveness of the WOMAC osteoarthritis index as compared with the SF-36 in patients with osteoarthritis of the legs undergoing a comprehensive rehabilitation intervention. Ann Rheum Dis 2001;60:834-40.
12. Baraliakos X, Hermann KGA, Landewe R, et al. Assessment of acute spinal inflammation in patients with ankylosing spondylitis by magnetic resonance imaging (MRI): A systematic comparison between contrast enhanced T1 and short-tau inversion recovery (STIR) sequences. Ann Rheum Dis 2005;64:1141-4.
13. Rahmouni A, Divine M, Mathieu D, et al. Detection of multiple myeloma involving the spine: efficacy of fat-suppression and contrast-enhanced MR imaging. AJR Am J Roentgenol 1993;160:1049-52.
14. Baur A, Stäbler A, Steinborn M, et al. Magnetic resonance tomography in plasmacytoma: ranking of various sequences in diffuse and focal infiltration patterns. Rofo Fortschr Geb Rontgenstr Neuen Bildgeb Verfahr 1998;168:323-9.
15. Stabler A, Krimmel K, Seiderer M, Gartner C, Fritsch S, Raum W. The nuclear magnetic resonance tomographic differentiation of osteoporotic and tumor-related vertebral fractures. The value of subtractive TR gradient-echo sequences, STIR sequences and Gd-DTPA. Rofo Fortschr Geb Rontgenstr Neuen Bildgeb Verfahr 1992;157:215-21.
16. Jones KM, Schwartz RB, Mantello MT, et al. Fast spin-echo MR in the detection of vertebral metastases: comparison of three sequences. AJNR Am J Neuroradiol 1994;15:401-7.