Low diagnostic accuracy and inter-observer agreement on CT and MRI in diagnosis of spinal fractures in multiple myeloma

Viktor Dalen,[1] Anne-Sofie Vegsgaard Olsen,[1] Claude-Pierre Jerome,[2] Jonn-Terje Geitung,[2,3] Anders E.A. Dahm[3,4]

[1]Medical Faculty, University of Oslo; [2]Department of Radiology, Akershus University Hospital, Lørenskog; [3]Institute of Clinical Medicine, University of Oslo; [4]Department of Haematology, Akershus University Hospital, Lørenskog, Norway

Abstract

Skeletal disease is common in multiple myeloma. We investigated the inter-observer agreement and diagnostic accuracy of spinal fractures diagnosed by computer tomography (CT) and magnetic resonance imaging (MRI) from 12 myeloma patients. Two radiologists independently assessed the images. CT, MRI, and other images were combined to a gold standard. The inter-observer agreement was assessed with Cohen's kappa. Radiologist 1 diagnosed 20 malignant spinal fractures on CT and 26 on MRI, while radiologist 2 diagnosed 12 malignant spinal fractures on CT and 22 on MRI. In comparison the gold standard diagnosed 10 malignant spinal fractures. The sensitivity for malignant fractures varied from 0.5 to 1 for CT and MRI, and the specificity varied from 0.17 to 0.67. On MRI, the specificity for malignant spinal fractures was 0.17 for both radiologists. The inter-observer agreement for malignant spinal fractures on CT was -0.42 (Cohen's kappa) and -0.13 for MRI, while for osteoporotic fractures it was -0.24 for CT and 0.53 for MRI. We conclude that malignant spinal fractures were over-diagnosed on CT and MRI. The inter-observer agreement was extremely poor.

Introduction

Multiple myeloma is a plasma cell neoplasia which accounts for 1% of all cancers and 10% of all haematological malignancies. The incidence in Europe is 4.5-6.0/100 000/year.[1] Multiple myeloma is diagnosed by detecting 10% or more plasma cells in the bone marrow. It is considered symptomatic if the patient has hypercalcaemia, renal insufficiency, anaemia, or bone lesions on skeletal radiography, computed tomography (CT) or positron emission tomography (PET) (the CRAB criteria). Moreover, plasma cell percentage in the bone marrow above 60%, free light chain ratio above 100, or 1 focal lesion on magnetic resonance imaging (MRI) is also considered as myeloma defining events since the risk of development of symptomatic multiple myeloma is high.[2] Symptomatic multiple myeloma usually means that treatment is initiated if the patient tolerates it. If it is asymptomatic, the recommended strategy is to watch and wait.[1]

Approximately 80-90% of multiple myeloma patients develop skeletal disease.[3] Screening for skeletal disease is therefore part of the standard diagnostic work-up in multiple myeloma. The preferred method was previously conventional x-ray of the skeleton. From 2015 low-dose CT, conventional-dose CT or PET-CT has been the recommended screening investigations for skeletal disease, if available.[2,4,5] The recommendation of low-dose CT was partly based on a systematic review from 2013,[6] which compared modern imaging methods including MRI, FDG-PET, PET-CT, and whole-body CT with conventional whole-body skeletal radiography. The review concluded that the newer imaging techniques were more sensitive than whole-body skeletal radiography and could detect up 80% more lesions. An important limitation of the evaluation of diagnostic imaging of skeletal disease in that systematic review was the lack of a diagnostic gold standard. In general, CT and MRI are assumed better than conventional x-ray based on general knowledge about the imaging techniques and because more bone lesions apparently are being observed on MRI and CT than on conventional x-ray. Without a diagnostic gold stander, this could hypothetically lead to false positive findings of skeletal disease in multiple myeloma patients.

Another problem, which is rarely investigated, is the variation in interpretation between radiologists. In the current study we aimed to find the inter-observer agreement and the diagnostic accuracy of spinal fractures in multiple myeloma patients who had taken CT and MRI.

Materials and methods

Correspondence: Anders Erik Astrup Dahm, Department of Haematology, Akershus University Hospital, Postbox 1000, 1478 Lørenskog, Norway.
Tel.: +47 67969330 - Fax: +47 67963820.
E-mail: aeadahm@gmail.com

Key words: Diagnostic specificity; malignant fractures; osteoporotic fractures; diagnostic imaging; haematology; multiple myeloma.

Contributions: AEAD, JTG and CPJ designed the study. VD and ASVGO collected the data. JTG and CPJ reviewed the images. AEAD, VD and ASVGO analysed the data and wrote the first draft. JTG and CPJ reviewed and edited the article. VD and ASVO contributed equally.

Conflict of interest: Dr. Dahm reports grants and personal fees from Pfizer AS, personal fees from Bristol-Mayer Squibb, Novartis Norway AS, and Bayer outside the submitted work. The other authors do not report any conflicts of interest.

Ethics approval and consent to participate: The project was approved by the local ethical committee at Akershus University Hospital (reference number 16-031), which decided that informed consent was not necessary.

Received for publication: 4 December 2020.
Accepted for publication: 10 September 2021.

Hematology Reports 2021; 13:9037
doi:10.4081/hr.2021.9037

Subjects

We searched the electronic patient journal for patients who had received a diagnosis with the ICD 10 code C 90.0, C 90.1, C 90.2, C 90.3 or D47 during the period from January 1st 2007 to December 31st 2015. We initially found 503 patients. We then restricted the search to those who had done MRI and CT within 4 months. We excluded patients, who did not have multiple myeloma, all patients without a fracture diagnosis, and all patients who did not have images of the entire spine for both CT and MRI. Finally, there were 12 CT scan images and 12 MRI images of the spine from 12 patients that could be further evaluated.

CT and MRI protocols

The CT data was collected on a Phillips CT iCT 256, kV 120 and mAs 30-50, rotation time 0.5 (low-dose CT) or kV 120 and mAs 250-300, rotation time 0.5 (conventional-dose CT). The MRI was a short protocol with sagittal slices, standard T1W and a standard and STIR T2W using a Phillips Ingenia 1.5T machine. All acquisitions on CT and MRI were performed without using contrast agents.

Evaluation of images

Two experienced radiologists (more than 30 years of experience) evaluated the images independently of previous descriptions and of each other. The images were evaluated with regard to malignant fractures and osteoporotic fractures. Radiologist 1 evaluated the CT images first and one month later the MRI images, while Radiologist 2 evaluated the MRI images first and then the CT images after one month. Five months after the initial evaluations both the radiologists sat together and used all the information available, previous images and descriptions, and made a consensus for the CT images and the MRI images and combined them to a diagnostic gold standard.

Statistics

The radiologists counted the number of fractures per patient for each imaging mode. For the inter-observer agreement between the two radiologists, the patients were categorized by whether or not they had any fracture or no fracture (yes or no). Inter-observer agreement was evaluated by cohen's kappa.

Results

Demographics

Eight of the 12 patients were male. The median age at the time of first image was 68.5 years (range 50-90). The median time from diagnosis to oldest image was 0 months (range 0-113). The median number of days between CT and MRI was 29.5 (range 0-98). Four of the 12 CT scans were low-dose CT.

Diagnostics of fractures

The number of malignant fractures found on CT and MRI compared with the gold standard is given in Table 1.

On CT the two radiologists had 24 diverging evaluations regarding the number of malignant fractures, only for one patient (patient 7) did they agree on the number of fractures. On MRI it was 18 diverging evaluations regarding malignant fractures between the two radiologists. Both radiologists diagnosed more malignant fractures separately on CT scan as well as on MRI as compared with the consensus and the gold standard. All the malignant fractures seen on the CT consensus were also seen on the MRI consensus, but more fractures were detected by MRI consensus. Table 2 shows the sensitivity, specificity, negative predictive value and positive predictive value for detection of malignant fracture on CT and MRI. As the Table 2 shows, the diagnostic accuracy, in particular the specificity of MRI, was poor.

The radiologist's evaluation of osteoporotic fractures is given in Table 3. On CT the two radiologists both diagnosed less osteoporotic fractures than the consensus and the gold standard. On MRI, however, the radiologists diagnosed more fractures than on the consensus and the gold standard. On CT the radiologists had 14 diverging evaluations, where eight were from the low-dose CT scans. On MRI the radiologists had 17 diverging judgements. In two patients, it was discovered osteoporotic fractures on CT that was not observed on

Table 1. Number of malignant fractures found on CT and MRI for the two radiologists.

Patient		CT			MRI		Gold standard
	R1	R2	Consensus	R1	R2	Consensus	
1	0	2	0	1	2	0	0
2	1	0	0	0	1	0	0
3	1	0	0	1	4	1	1
4*	0	3	0	0	1	2	2
5	2	1	0	1	2	2	2
6	0	2	0	1	1	0	0
7	0	0	0	6	3	0	0
8	8	0	1	9	4	1	1
9	1	0	0	1	0	0	0
10*	3	4	0	2	1	1	1
11*	1	0	0	1	1	0	0
12*	3	0	3	3	2	3	3
Total	20	12	4	26	22	10	10
Diverging evaluations	24		18				

* Denotes low dose CT. R1, radiologist 1; R2, radiologist 2.

Table 2. Sensitivity and specificity for detection of malignant fractures on CT and MRI.

	CT		MRI	
	R1	R2	R1	R2
Sensitivity	0.83	0.5	0.83	1
Specificity	0.5	0.67	0.17	0.17
Negative predictive value	0.63	0.6	0.50	0.55
Positive predictive value	0.75	0.57	0.50	1

R1, radiologist 1; R2, radiologist 2.

MRI. The sensitivity, specificity, positive predictive value and negative predictive value for diagnosis of osteoporotic fractures was poor (Table 4). Although all the patients were selected based on a previous diagnosis of fracture, four of the 12 patients did in fact not have any fracture, neither malignant, nor osteoporotic.

Inter-observer agreement of fracture diagnosis

When the data was categorized by whether the patients had fractures or not, we found that the radiologists agreed less than one would expect by chance alone. Thus, the Cohen's kappa became negative for the diagnosis of both malignant and osteoporotic fractures on CT, and negative for malignant fracture on MRI. Only the diagnosis of osteoporotic fracture on MRI showed a positive kappa of 0.53 (Table 5). It appeared that the radiologists disagreed more on CT than on MRI.

Discussion

The two main results of the current study was the extremely low agreement between the two radiologists in the diagnosis of spinal fractures in multiple myeloma patients, and the low diagnostic accuracy of CT and MRI. The low agreement was evident for both CT and MRI, while the low diagnostic accuracy in particular resulted in too many diagnosed malignant fractures on both MRI and CT, resulting in very low specificity, particularly for MRI. For osteoporotic fractures the overdiagnosis was confined to the MRI modality.

MRI is considered the imaging gold standard for bone marrow involvement of myeloma, but not for bone destruction. It is also considered the best method to separate malignant fractures from osteoporotic fractures.[7] A previous study have shown that MRI diagnose more bone lesions in multiple myeloma than CT, while in another study ultra-low dose CT diagnosed more axial lesions than MRI.[8,9] In our study, both radiologists diagnosed more malignant and osteoporotic fractures on MRI than CT, but compared with the gold standard the two radiologists grossly overestimated the number of especially malignant fractures with both methods. Since CT is better in evaluating the bone cortex, but MRI better to see malignant infiltration, a possible explanation is that MRI and CT complement each other so it is easier to separate osteoporotic and malignant fractures when both image modalities are considered at the same time. Thus, a possible gold standard in future studies could be the combination of conventional dose CT and MRI.

Our study suggests that more sensitive methods for discovering myeloma bone disease may result in more false positive findings. The ideal gold standard in myeloma skeletal disease would be to take biopsy from each of the fractures to determine malignancy, but that is not feasible since it

Table 3. Number of ostoeporotic fractures found on CT and MRI for the two radiologists

Patient	CT			MRI			Gold standard
	R1	R2	Consensus	R1	R2	Consensus	
1	3	0	2	3	1	2	2
2	2	2	3	10	8	4	4
3	0	1	0	0	0	0	0
4*	3	0	0	6	2	0	0
5	0	1	2	0	0	0	1
6	1	0	0	0	1	0	0
7	0	0	0	0	0	0	0
8	0	0	5	0	7	4	5
9	0	0	0	0	1	0	0
10*	0	0	2	3	3	2	2
11*	0	1	0	0	0	0	0
12*	0	4	0	1	1	0	0
Total	9	9	14	23	24	12	14
Diverging evaluations	14		17				

* Denotes low dose CT. R1, radiologist 1; R2, radiologist 2.

Table 4. Sensitivity and specificity for detection of osteoporotic fractures on CT and MRI.

	CT		MRI	
	R1	R2	R1	R2
Sensitivity	0.40	0.40	0.60	0.8
Specificity	0.71	0.57	0.71	0.43
Negative predictive value	0.50	0.40	0.60	0.50
Positive predictive value	0.63	0.57	0.71	0.75

R1, radiologist 1; R2, radiologist 2.

Table 5. Inter-observer agreement of spinal fracture diagnosis.

Investigation	CT malignant fractures	CT osteoporotic fractures	MRI malignant fractures	MRI osteoporotic fractures
Disagreement	9 of 12	7 of 12	3 of 12	3 of 12
Cohen's kappa	-0.42	-0.24	-0.13	0.53

would mean several biopsies from the spine in each patient. A gold standard is, nevertheless, a consistent problem in previous studies of myeloma bone disease. For example, in the systematic review of imaging techniques in multiple myeloma by Regelink et al[6] conventional whole-body x-ray is used as the reference method. Based on this study, it is stated that CT and MRI are more sensitive than conventional radiography because more lesions can be seen. This is an imprecise use of the term "sensitivity" given the lack of a diagnostic gold standard in all the studies reviewed. It is in fact clear from that review and several other studies that suspected malignant lesions can be seen on whole body x-ray or on CT that is not seen on MRI and vice versa.[10-12]

The current article illustrates that CT and MRI may wrongly diagnose skeletal disease in myeloma. Findings of skeletal disease in multiple myeloma can be decisive for whether treatment is started or not. In one study, of 138 patients of early multiple myeloma, bone involvement was the only CRAB criterion in 40 patients.[13] Since guidelines state that most asymptomatic myeloma patients should not be treated,[1] a wrong diagnosis of skeletal disease in a myeloma patient may result in unnecessary treatment.

It is common to use Cohen's kappa to assess inter-agreement between radiologists. Which level of kappa that is considered good enough is open for interpretation, but kappa >0.80 is considered as very good agreement, while kappa <0.20 is considered as almost no agreement.[14] Not many studies have investigated inter-observer agreement of fractures in multiple myeloma. Zacchino and co-workers found variable inter-agreement between four readers in a study of 100 multiple myeloma patients evaluating both osteolytic lesions and fractures with whole body low-dose CT. Regarding fractures in the spine they found a kappa varying from approximately 0.2 to 0.5, with lowest agreement for fractures in the cervical spine.[15] Thus, Zacchion's study showed low agreement, but not as low as in our study. Of note, in that study, the agreement of osteolytic lesions was generally higher than for fractures. In another recent study inter-observer agreement was evaluated for CT and MRI in 22 myeloma patients using four reviewers. They assessed agreement with intraclass correlation coefficients and found low inter-observer agreement for CT, but high agreement for MRI for presence of myeloma skeleton disease.[16] Also in our study we found better agreement for MRI than for CT, but the agreement was generally almost nonexistent. Thus, our study, along with other studies, show that low inter-observer agreement in interpretation of myeloma skeleton disease on CT is a problem. Our study suggests that it may be a problem for evaluation of fractures in the columnal also on MRI. This means that the diagnosis of myeloma bone disease partly depends on the radiologist evaluating the images.

The main weakness of the current study is the low number of patients. The reason was that not many patients had taken CT and MRI within 4 months. Another possible weakness is that the patients could have acquired new skeletal lesions during the 4 months between the investigations, although the median number of days between CT and MRI was 29.5 days. An important limitation is also that we focused only on spinal fractures. What is clinically important in multiple myeloma is to diagnose any skeletal lesions caused by myeloma, but multiple myeloma patients can have osteoporotic or malignant fractures, sometimes sclerotic skeletal disease, and osteoporotic bone disease. The restriction to fractures in our study was a pragmatic decision to focus on one type of skeletal disease in to minimize the variation between the evaluations of the radiologists. The use of a consensus of CT and MRI as a gold standard can be considered a weakness, and may not be a true gold standard, but is probably better than no gold standard which has been the general rule in other studies.[6] A strength of the study is that agreement between radiologists in the interpretation of multiple myeloma skeletal disease is a problem that is rarely addressed. Another strength of the current study is that, to our knowledge, we are the first to try to make a diagnostic gold standard for malignant fractures in multiple myeloma.

Conclusions

Over-diagnosis of malignant fractures and high variability between radiologist interpretations may result in wrong diagnosis of spinal fractures in multiple myeloma-patients. We suggest that new studies on diagnostic imaging use a combination of full-dose CT and MRI as a diagnostic gold standard.

References

1. Moreau P, San Miguel J, Sonneveld P, et al. Multiple myeloma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2017;28:iv52-61.
2. Rajkumar SV, Dimopoulos MA, Palumbo A, et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. Lancet Oncol 2014;15:e538-48.
3. Kyle RA, Gertz MA, Witzig TE, et al. Review of 1027 patients with newly diagnosed multiple myeloma. Mayo Clin Proc 2003;78:21-33.
4. Chantry A, Kazmi M, Barrington S, et al. Guidelines for the use of imaging in the management of patients with myeloma. Br J Haematol 2017;178:380-93.
5. Hillengass J, Usmani S, Rajkumar SV, et al. International myeloma working group consensus recommendations on imaging in monoclonal plasma cell disorders. Lancet Oncol 2019;20:e302-12.
6. Regelink JC, Minnema MC, Terpos E, et al. Comparison of modern and conventional imaging techniques in establishing multiple myeloma-related bone disease: a systematic review. Br J Haematol 2013;162:50-61.
7. Dimopoulos MA, Hillengass J, Usmani S, et al. Role of magnetic resonance imaging in the management of patients with multiple myeloma: a consensus statement. J Clin Oncol 2015;33:657-64.
8. Baur-Melnyk A, Buhmann S, Becker C, et al. Whole-body MRI versus whole-body MDCT for staging of multiple myeloma. AJR Am J Roentgenol 2008;190:1097-104.
9. Ippolito D, Talei Franzesi C, Spiga S, et al. Diagnostic value of whole-body ultra-low dose computed tomography in comparison with spinal magnetic resonance imaging in the assessment of disease in multiple myeloma. Br J Haematol 2017;177:395-403.
10. Mahnken AH, Wildberger JE, Gehbauer G, et al. Multidetector CT of the spine in multiple myeloma: comparison with MR imaging and radiography. AJR Am J Roentgenol 2002;178:1429-36.
11. Gleeson TG, Moriarty J, Shortt CP, et al. Accuracy of whole-body low-dose multidetector CT (WBLDCT) versus skeletal survey in the detection of myelomatous lesions, and correlation of disease distribution with whole-body MRI (WBMRI). Skeletal Radiol 2009;38:225-36.
12. Hillengass J, Moulopoulos LA, Delorme S, et al. Whole-body computed tomography versus conventional skeletal survey in patients with multiple myeloma: a study of the International

Myeloma Working Group. Blood Cancer J 2017;7:e599.

13. Ippolito D, Besostri V, Bonaffini PA, et al. Diagnostic value of whole-body low-dose computed tomography (WBLDCT) in bone lesions detection in patients with multiple myeloma (MM). Eur J Radiol 2013;82:2322-7.
14. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
15. Zacchino M, Bonaffini PA, Corso A, et al. Inter-observer agreement for the evaluation of bone involvement on Whole Body Low Dose Computed Tomography (WBLDCT) in Multiple Myeloma (MM). Eur Radiol 2015;25:3382-9.
16. Lai AYT, Riddell A, Barwick T, et al. Interobserver agreement of whole-body magnetic resonance imaging is superior to whole-body computed tomography for assessing disease burden in patients with multiple myeloma. Eur Radiol 2020;30:320-7.