ORIGINAL RESEARCH

Metastatic Versus Osteoporotic Vertebral Fractures on MRI: A Blinded, Multicenter, and Multispecialty Observer Agreement Evaluation

Estanislao Arana, MD, MHE, PhD[a,b]; Francisco M. Kovacs, MD, PhD[b,c]; Ana Royuela, PhD[b,d];
Beatriz Asenjo, MD, PhD[b,e]; Fatima Nagib, MD[b,e]; Sandra Pérez-Aguilera, MD[b,f];
María Dejoz, BEng[b,g]; Alberto Cabrera-Zubizarreta, MD[b,h]; Yolanda García-Hidalgo, MD, PhD[b,i];
and Ana Estremera, MD, PhD[b,j]; for the Spanish Back Pain Research Network Task Force for
the Improvement of Inter-Disciplinary Management of Spinal Metastasis*

ABSTRACT

Background: MRI is assumed to be valid for distinguishing metastatic vertebral fractures (MVFs) from osteoporotic vertebral fractures (OVFs). This study assessed (1) concordance between the image-based diagnosis of MVF versus OVF and the reference (biopsy or follow-up of >6 months), (2) interobserver and intraobserver agreement on key imaging findings and the diagnosis of MVF versus OVF, and (3) whether disclosing a patient's history of cancer leads to variations in diagnosis, concordance, or agreement. **Patients and Methods:** This retrospective cohort study included clinical data and imaging from 203 patients with confirmed MVF or OVF provided to 25 clinicians (neurosurgeons, radiologists, orthopedic surgeons, and radiation oncologists). From January 2018 through October 2018, the clinicians interpreted images in conditions as close as possible to routine practice. Each specialist assessed data twice, with a minimum 6-week interval, blinded to assessments made by other clinicians and to their own previous assessments. The kappa statistic was used to assess interobserver and intraobserver agreement on key imaging findings, diagnosis (MVF vs OVF), and concordance with the reference. Subgroup analyses were based on clinicians' specialty, years of experience, and complexity of the hospital where they worked. **Results:** For diagnosis of MVF versus OVF, interobserver agreement was fair, whereas intraobserver agreement was substantial. Only the latter improved to almost perfect when a patient's history of cancer was disclosed. Interobserver agreement for key imaging findings was fair or moderate, whereas intraobserver agreement on key imaging findings was moderate or substantial. Concordance between the diagnosis of MVF versus OVF and the reference was moderate. Results were similar regardless of clinicians' specialty, experience, and hospital category. **Conclusions:** When MRI is used to distinguish MVF versus OVF, interobserver agreement and concordance with the reference were moderate. These results cast doubt on the reliability of basing such a diagnosis on MRI in routine practice.

J Natl Compr Canc Netw 2020;18(3):267–273
doi: 10.6004/jnccn.2019.7367

Background

Nontraumatic vertebral fractures are frequently seen in clinical practice. Most are caused by osteoporosis and are diagnosed as osteoporotic vertebral fractures (OVFs), but metastatic vertebral fractures (MVFs) are also common. Determining whether a vertebral fracture has been caused by MVF or OVF is key for establishing appropriate treatment and prognosis and can have a profound psychological impact on patients. Therefore, the accuracy and reliability of the data used to reach this diagnosis are paramount.

Several imaging findings are frequently used to help distinguish between OVF and MVF.[1,2] Some have been fed into risk-scoring algorithms developed to identify patients at a higher risk of experiencing MVF.[3] To be useful in clinical practice and lead to sound treatment decisions, risk-assessment algorithms should be evidence-based and built on parameters that can be assessed reliably. However, the available risk-scoring algorithms in this field rely on ancillary imaging findings, for which

[a]Department of Radiology, Fundación Instituto Valenciano de Oncología, Valencia; [b]Spanish Back Pain Research Network, Kovacs Foundation, Palma de Mallorca; [c]Unidad de la Espalda Kovacs, Hospital Universitario HLA-Moncloa, Madrid; [d]Clinical Biostatistics Unit, Instituto de Investigación Sanitaria Puerta de Hierro-Segovia de Arana, Madrid; CIBERESP; [e]Department of Radiology, Hospital Universitario Regional de Málaga, Málaga; [f]Department of Radiology, Hospital de Manacor, Mallorca; [g]School of Biomedical Engineering, Universitat Politècnica de Valencia, Valencia; [h]Department of Radiology, Hospital de Galdakao, Galdakao, Bizkaia; [i]Department of Radiology, Hospital Universitario Puerta de Hierro, Madrid; and [j]Department of Radiology, Hospital Son Llàtzer, Palma de Mallorca, Spain.

*To view additional members of the Spanish Back Pain Research Network Task Force, see supplemental eAppendix 1 (available with this article at JNCCN.org).

See JNCCN.org for supplemental online content.

See page 362 for related commentary.

reliability is unknown. The need to assess their reliability has been previously highlighted.[3,4]

The available scoring systems have been developed based on the interpretation of images by only one observer[1] or on the consensus of readers working in the same institution who tested the validity of their scoring systems with a small number of patients.[2,3] However, in clinical practice, when patients seek care for back pain caused by a nontraumatic vertebral fracture, spine imaging can be assessed by practitioners from an array of specialties, and management of OVF and especially MVF is multidisciplinary.

Therefore, the purpose of this study was to assess among clinicians from different specialties and working in different healthcare centers, in conditions as close as possible to routine clinical practice, (1) concordance between the clinical diagnosis (MVF vs OVF) and the reference (diagnosis established by biopsy or clinical follow-up), (2) interobserver and intraobserver agreement on the diagnosis of MVF versus OVF and on the interpretation of key imaging findings leading to such diagnosis, and (3) whether concordance and agreements improve when clinicians are aware of a patient's history of cancer.

Patients and Methods

This study was approved by the Institutional Review Boards of the participating hospitals and complied with the Guidelines for Reporting Reliability and Agreement Studies.[5] Written informed consent was waived because of the retrospective nature of this study.

Setting and Participants

Patients and images were selected by a radiologist with 25 years of experience who did not participate in image interpretation. He revised records from his hospital in reverse chronologic order (ie, more recent cases were revised first) and selected cases complying with the inclusion criteria until the sample size was reached. The radiologist then selected 3 images per patient: 2 sagittal images on T1-, T2-, or short inversion time inversion-recovery (STIR)–weighted images, and 1 axial T1-weighted image.

Inclusion criteria were having requested care for a nontraumatic vertebral fracture, and diagnosis of MVF or OVF confirmed through biopsy or clinical follow-up of >6 months. Exclusion criteria were missing clinical history for any of the data required by the readers, and imaging of insufficient quality to assess the spinal levels affected (Figure 1).

A total of 22 hospital departments of radiology, radiation oncology, orthopedic surgery, and neurosurgery were invited to join the study because they had participated in previous spine studies undertaken by the Spanish Back Pain Research Network or had expressed interest in doing so. The hospital departments were located in 18 hospitals across 12 geographic regions; 6 departments were located in 5 private hospitals and 16 were located in 13 nonprofit hospitals belonging to or working for the Spanish National Health System (SNHS). The SNHS is the tax-funded, government-owned organization that provides free healthcare to every resident in Spain. The SNHS classifies hospitals into 5 categories based on their complexity,[6] with category 1 the simplest and category 5 the most complex. Departments invited to participate in this study were located in category 2 through 5 hospitals.

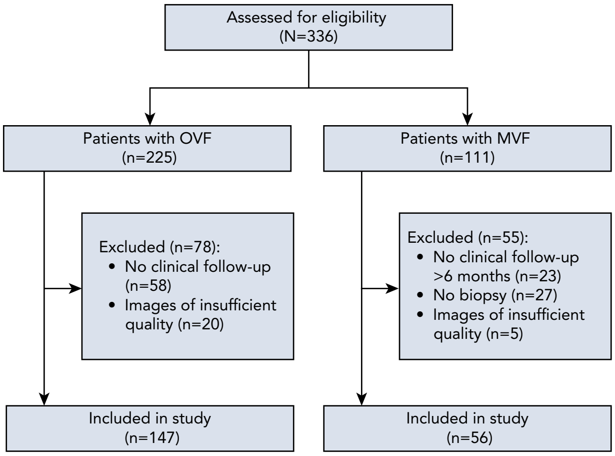


Figure 1. Flowchart of the selection process.
Abbreviations: MVF, metastatic vertebral fracture; OVF, osteoporotic vertebral fracture.

According to standard procedure in our setting, neither subjects nor clinicians received any compensation for their involvement in this study.

MRI Evaluation, Reporting, and Interpretation

All images were acquired on 4 1.5T MRI systems, using similar sequences (supplemental eTable 1, available with this article at JNCCN.org).

The recruiting radiologist prepared an information pack on each patient containing 3 images and a clinical vignette summarizing the patient's age, oncologic history, and clinical signs and symptoms.[7] Patient identity was masked and a code was assigned to each pack. All packs were uploaded to an online platform designed for this study (http://www.typeform.com/). The 3 images included 2 sagittal images on T1-, T2-, or STIR-weighted images and 1 axial T1-weighted image. The radiologist segmented the selected images so that readers were shown the index vertebral segment, the one immediately above, and the one immediately below. In the case of patients showing vertebral fractures at several levels, the radiologist defined the index as the one showing a recent fracture, at the level for which the patient had requested care, and that was subject to biopsy or clinical follow-up for >6 months.

MRI findings assessed in this study were selected through a literature review[1,2] and are shown in supplemental eTable 2. They include those findings used to calculate the MRI Evaluation Totalizing Assessment (META) score.[1] The readers assessed all MRI images on their own, prospectively, from January 2018 through October 2018, using an in-house online MRI interpretation system. No attempt was made to homogenize their diagnostic criteria or interpretation of images. Readers were told to use their own clinical judgment as they would in routine clinical practice and to upload the report directly onto the online platform. After they assessed the imaging findings, readers were requested to state their diagnosis ("MVF" vs "OVF"). Finally, after the patient's cancer history was disclosed, readers were given the opportunity to modify their diagnosis (Figure 2), and modifications were recorded.

Readers assessed each information pack twice, with a minimum 6-week interval between the 2 rounds. After the information from the first round was uploaded, the platform software made it impossible for readers to access it again until the interval had elapsed. It also denied access to colleagues' reports and to their own previous reports.

Data introduced into the platform were automatically converted into a spreadsheet. The software engineer in charge of developing the platform cross-checked to ensure that data in the database matched the information that readers had introduced into the platform.

Statistical Analysis

To assess interobserver and intraobserver agreement, ratings from each observer were cross-tabulated, and agreement was measured using the kappa statistic (κ) with the corresponding 95% confidence interval for interobserver agreement and the percentiles 25 and 75 (interquartile range [IQR], p25–p75) for intraobserver agreement. Kappa values were categorized as reflecting an "almost perfect" (0.81–1.00), "substantial" (0.61–0.80), "moderate" (0.41–0.60), "fair" (0.21–0.40), "slight" (0.00–0.20), or "poor" (<0.00) agreement.[8]

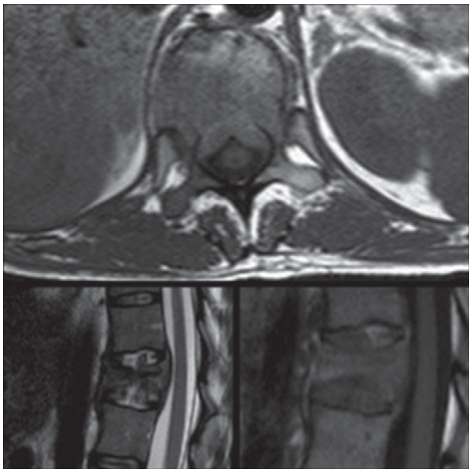
The association between the diagnosis (MVF vs OVF) before and after readers were aware of a patient's cancer history was based on the assessments made during the first round. Diagnostic accuracy was defined as the concordance between each reader's diagnosis at the first round (MVF vs OVF) and the reference diagnosis. Concordance was measured using the kappa statistic. In a subgroup analysis, diagnostic accuracy was measured separately for subjects who presented and did not present previous fractures on imaging.

Sample size was estimated at 203 patients with vertebral fractures, assuming that (1) vertebral fractures would be caused by MVF in 25% to 30% of cases,[9] (2) the minimal number of assessments to be compared would be 2 (for intraobserver agreement), and (3) the kappa index would be ≥0.7 with a confidence margin of 0.10 on each side.

Results

All 22 hospital departments invited to join the study accepted, and 25 clinicians from these departments participated: 9 radiologists, 4 radiation oncologists, 5 orthopedic surgeons, and 7 neurosurgeons (Table 1). The number of years (after residency) that the clinicians had been interpreting spine MRIs in routine practice on a daily basis ranged from 4 to 35 years. Table 1 also shows the characteristics of the 203 patients whose clinical histories and images were selected for the study and of the 25 readers who interpreted their data.

Case 1. Female, age 34 y



- Normal vertebral signal replace with bone marrow edema
  - □ Pratially or completely
  - □ Showing a bandlike pattern
- Horizontal frecture line on fluid-sensitive sequence (STIR) or T2-weighted imaging: □ Yes / □ No
- Deposit-like appearance of pedicle involvement: □ Yes / □ No
- Convexity of the posterior vertebral body border (bulging posterior cortex): □ Yes / □ No
- Posterosuperior retropulsion: □ Yes / □ No
- Symmetry of the signal intensity changes: □ Symmetric □ Asymmetric
- Diagnosis: □ VFO □ MVF
- This patient has a history of cancer: □ Yes / □ No
- Do you want to modify your diagnosis? □ Yes / □ No
- Diagnosis: □ OVF □ MVF

Figure 2. Sample imaging finding.
Abbreviations: MVF, metastatic vertebral fracture; OVF, osteoporotic vertebral fracture; STIR, short inversion time inversion-recovery.

Table 1. Sample Characteristics

Characteristic	n (%)
Patients	
Total, N	203
Age, mean (SD), y	60.8 (12.3)
Women	62.1 (14.5)
Men	61.7 (10.8)
Sex (female)	139 (68.47)
Location of spinal fracture	
Thoracic	98 (48.27)
Lumbar	105 (51.73)
Diagnosis[a]	
OVF	147 (72.4)
MVF	56 (27.6)
History of cancer	
No	122 (60.1)
Yes	81 (39.9)
Previous spine fractures	
No	132 (65)
Yes	71 (35)
Location of previous spine fracture	
Thoracic	36 (50.7)
Lumbar	35 (49.3)
Primary malignancies	
Lung	20 (35.7)
Breast	16 (28.5)
Colon	8 (14.3)
Lymphoma or myeloma	4 (7.1)
Other	8 (14.3)
Readers	
Total, N	25
Specialty	
Radiology	9 (36.0)
Radiation oncology	4 (16.0)
Orthopedic surgery	5 (20.0)
Neurosurgery	7 (28.0)
Years of experience (postresidency), y	
≤7	7 (32.5)
8–13	7 (30.1)
≥14	11 (37.4)
Hospital category (complexity)[b]	
2	3 (3.6)
3	7 (30.1)
4	7 (22.9)
5 (most complex)	8 (43.4)
Hospitals	
Total, N	18
Management	
Nonprofit[c]	13 (72.2)
For-profit[d]	5 (27.8)
Departments, N	22
Radiology	8 (40.0)
Radiation oncology	4 (20.0)
Orthopedic surgery	3 (15.0)
Neurosurgery	7 (25.0)

(continued)

Abbreviations: MVF, metastatic vertebral fracture; OVF, osteoporotic vertebral fracture; SNHS, Spanish National Health System.
[a] Diagnosis established by the reference (biopsy or follow-up >6 months).
[b] Based on size, availability of high-tech medical equipment and procedures, and degree of educational activity. No readers from category 1 hospitals (simplest) were included in this study.
[c] Belonging to and managed by the SNHS, or belonging to or managed by charities working for the SNHS.
[d] Privately owned and managed.

As Table 2 shows, interobserver agreement in the diagnosis of MVF versus OVF was fair (κ, 0.397; 95% CI, 0.347–0.450) when the reader was unaware of the patient's history of cancer. When the patient's history of cancer was disclosed, the agreement increased to moderate (κ, 0.467; 95% CI, 0.418–0.518).

Intraobserver agreement on the diagnosis of MVF versus OVF was substantial (κ, 0.624; IQR, 0.517–0.693), and improved to almost perfect after the patient's history of cancer was disclosed (κ, 0.878; IQR, 0.781–0.939 and κ, 0.851; IQR, 0.779–0.948 at the first and second rounds, respectively). This increase in agreement was observed across all clinical specialties, with orthopedic surgery showing the highest increase (from κ, 0.588; IQR, 0.509–0.595 to κ, 0.917; IQR, 0.859–0.959) (Table 3).

Interobserver agreement was moderate on "deposit-like appearance of pedicle involvement" and "bulging posterior cortex" and fair on all the other imaging findings (supplemental eTable 3). Agreement among radiologists was moderate for most imaging findings, but no consistent differences were found among clinical specialties (supplemental eTable 3).

Intraobserver agreement on individual imaging findings ranged from moderate to substantial and was similar across clinical specialties (supplemental eTable 4).

After being informed of a patient's clinical history of cancer, the readers modified the diagnosis (MVF vs OVF) of 142 patients (69.5%). All the readers modified the diagnosis of at least 1 patient (range of number of patients for whom each clinician changed the diagnosis, 1–39). Among the 5,075 assessments made by the 25 readers using the 203 images, the previous diagnosis was changed in

Table 2. Interobserver Agreement

	Kappa (95% CI)
All readers (n=25)	
Diagnosis of OVF vs MVF (before disclosing history of cancer)	0.397 (0.347–0.450)
Pattern of signal abnormalities	0.396 (0.349–0.445)
Horizontal fracture line	0.220 (0.177–0.266)
Deposit-like appearance of pedicle involvement	0.447 (0.395–0.501)
Bulging posterior cortex	0.426 (0.383–0.472)
Posterosuperior retropulsion	0.319 (0.280–0.359)
Symmetry of signal intensity changes	0.270 (0.230–0.312)
Diagnosis of OVF vs MVF (after disclosing history of cancer)	0.467 (0.418–0.518)
Radiology (n=9)	
Diagnosis of OVF vs MVF (before disclosing history of cancer)	0.508 (0.446–0.573)
Diagnosis of OVF vs MVF (after disclosing history of cancer)	0.574 (0.518–0.633)
Neurosurgery (n=7)	
Diagnosis of OVF vs MVF (before disclosing history of cancer)	0.364 (0.305–0.425)
Diagnosis of OVF vs MVF (after disclosing history of cancer)	0.456 (0.397–0.518)
Orthopedic surgery (n=5)	
Diagnosis of OVF vs MVF (before disclosing history of cancer)	0.342 (0.275–0.411)
Diagnosis of OVF vs MVF (after disclosing history of cancer)	0.370 (0.303–0.438)
Radiation oncology (n=4)	
Diagnosis of OVF vs MVF (before disclosing history of cancer)	0.321 (0.256–0.389)
Diagnosis of OVF vs MVF (after disclosing history of cancer)	0.394 (0.325–0.465)

Abbreviations: MVF, metastatic vertebral fracture; OVF, osteoporotic vertebral fracture.

5.0% of the patients without a history of cancer versus 10.8% of those with a history of cancer (chi-square, $P<.001$).

Before readers were aware of a patient's clinical history of cancer, concordance of their diagnosis with the reference was moderate (κ, 0.437; IQR, 0.326–0.511). Having access to a patient's history only marginally improved concordance (κ, 0.443; IQR, 0.398–0.526). Diagnostic accuracy was only fair for orthopedic surgeons, whereas it was moderate for all other specialties. However, differences in κ values were minimal, and the IQR values overlapped. Diagnostic accuracy was very similar regardless of years of professional experience and category of hospital (supplemental eTable 5).

Concordance with the reference for subjects without images of preexisting fractures was κ=0.452 (IQR, 0.387–0.509) before the clinical history of cancer was disclosed and κ=0.462 (IQR, 0.407–0.570) after it was disclosed. For subjects with preexisting fractures, these values were κ=0.286 (IQR, 0.183–0.396) and κ=0.331 (IQR, 0.219–0.368), respectively (supplemental eTable 6).

Discussion

In routine practice, the suspicion of MVF or OVF is based on clinical history and imaging. Our findings showed that interobserver agreement was fair and that diagnostic accuracy was moderate.

This is the first study to analyze the reliability of the diagnosis of MVF versus OVF using a large multidisciplinary team of readers working in different healthcare centers and assessing diagnostic accuracy against a reference. It was conducted in conditions as close as possible to routine clinical practice; readers were provided with actual clinical histories.[10] Because no instructions, scoring systems, or meetings were implemented to improve agreement,[11–13] clinicians had to make their diagnosis on their own based on data from clinical history and imaging, with common heuristics and biases.[14] All of these factors may account for differences between the results of this study and the almost perfect agreement reported by the medical professionals who developed the META score (κ, 0.93),[1] which previous studies have shown to not be reproducible.[15]

In this study, readers were experts who had been managing vertebral fractures and interpreting spine imaging for up to 35 years, had participated in previous research in this field, and felt confident enough to volunteer for a study assessing their interpretation of spine images. Diagnostic accuracy was very similar across clinical specialties, readers' experience, and hospital category and was consistent with results from the few previous studies that analyzed the reproducibility of single imaging findings and the META score.[1,15] Therefore, fair interobserver agreement and moderate diagnostic accuracy may be the best that can be realistically expected when using MRI to distinguish MVF versus OVF in routine practice, simply because with current technology, images of MVF and OVF are sometimes indistinguishable.[16,17] For instance, "bulging posterior cortex" was one of the imaging findings with the best interobserver agreement found in this and previous studies, and specifically, expansion of the posterior aspect of the vertebral contour is associated with malignant fractures.[18] However, it can also be observed in benign OVFs, especially in acute posttraumatic fractures.[16]

The low reproducibility of imaging findings challenges the validity of purportedly evidence-based decision support systems based on them.[2] In fact, a decision system based on unreliable findings can be detrimental.[3] The degree of agreement found in this and previous studies would classify MRI as class II for diagnosing MVF versus OVF and as class III for assessing individual imaging findings.[19]

In general, disclosing accurate clinical data slightly increases the accuracy of diagnostic tests.[20] For imaging

Table 3. Intraobserver Agreement

	Median Kappa (IQR)
All readers (n=25)	
Diagnosis of OVF vs MVF (agreement between diagnosis in both rounds, before disclosing history of cancer)[a]	0.624 (0.517–0.693)
Pattern of signal abnormalities	0.660 (0.555–0.762)
Horizontal fracture line	0.535 (0.457–0.683)
Deposit-like appearance of pedicle involvement	0.653 (0.549–0.732)
Bulging posterior cortex	0.715 (0.618–0.824)
Posterosuperior retropulsion	0.673 (0.592–0.731)
Symmetry of signal intensity changes	0.489 (0.402–0.646)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), first round[b]	0.878 (0.781–0.939)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), second round[c]	0.851 (0.779–0.948)
Radiology (n=9)	
Diagnosis of OVF vs MVF (agreement between diagnosis in both rounds, before disclosing history of cancer)[a]	0.652 (0.630–0.733)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), first round[b]	0.867 (0.805–0.881)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), second round[c]	0.845 (0.779–0.870)
Neurosurgery (n=7)	
Diagnosis of OVF vs MVF (agreement between diagnosis in both rounds, before disclosing history of cancer)[a]	0.550 (0.483–0.693)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), first round[b]	0.877 (0.713–0.979)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), second round[c]	0.948 (0.832–1.000)
Orthopedic surgery (n=5)	
Diagnosis of OVF vs MVF (agreement between diagnosis in both rounds, before disclosing history of cancer)[a]	0.588 (0.509–0.595)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), first round[b]	0.917 (0.859–0.959)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), second round[c]	0.880 (0.871–0.930)
Radiation oncology (n=4)	
Diagnosis of OVF vs MVF (agreement between diagnosis in both rounds, before disclosing history of cancer)[a]	0.618 (0.575–0.683)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), first round[b]	0.912 (0.706–0.958)
Diagnosis of OVF vs MVF (agreement before and after disclosing history of cancer), second round[c]	0.761 (0.581–0.921)

Abbreviations: IQR, interquartile range; MVF, metastatic vertebral fracture; OVF, osteoporotic vertebral fracture.

[a] This kappa value reflects the agreement between the diagnosis established by the same reader based on the same images, at the first and second rounds (in both cases, before being aware of patient's history of cancer).

[b] This kappa value reflects the agreement between the diagnosis established by the same reader based on the same images at the first round, before and after being aware of patient's cancer history.

[c] This kappa value reflects the agreement between the diagnosis established by the same reader based on the same images at the second round, before and after being aware of patient's cancer history.

procedures, some studies have suggested that accurate data disclosure decreases the interpretative performance,[21] whereas others have denied any negative consequences.[22] In our study, disclosing a patient's cancer history had no significant impact on interobserver agreement or diagnostic accuracy, but increased intraobserver agreement significantly and led to changes in the diagnosis of MVF versus OVF in 69.5% of the cases.

Diagnostic performance was similar across specialties. This is consistent with previous studies on the interpretation of spine imaging.[23–26] For patients with metastatic spine disease, surgeons' assessment of imaging is often considered the reference for referral to surgery.[27] However, no significant differences existed across surgical and nonsurgical specialties when these clinicians assessed the spinal instability score.[28] No patient was excluded due to sclerotic metastases, previous trauma history, or myeloma, in which vertebral signal intensity changes are misleading.[16,17] Interobserver agreement and diagnostic accuracy may be different for patients showing these findings.

This study has several limitations. The cases analyzed were selected by a radiologist and were not a random sample. These conditions were necessary to select a sample with the desired proportion of cases with MVF confirmed by a reference and is common practice in agreement studies on imaging or concordance.[13,26] In this study, readers only assessed 3 images, whereas in clinical practice physicians review multiple images. This rule was decided at the design phase of the study to enhance participation. Moreover, it is common practice in agreement studies to restrict the number of images to the most relevant or potentially confounding ones.[12,28] None of the selected cases showed findings highly suggestive of malignancy, such as soft tissue mass, which commonly lead to higher agreement between orthopedic surgeons and radiologists.[29] Therefore, it is possible that agreement would have been higher if a number of patients included in the study had shown these findings. However, this study aimed to assess agreement in conditions as close as possible to clinical practice, and inclusion criteria did not require any specific finding. The classification of imaging findings did not follow the categories established by the META score. This condition was decided at the design phase of the study because these categories have been shown to be unreliable.[15] Using different image sequences may lead to different results. However, MR imaging sequences are not widely available,[17] and were therefore considered inappropriate for a study replicating routine practice as closely as possible. Nevertheless, future studies should explore the impact of different image sequences on agreement and diagnostic accuracy.[17]

Conclusions

Diagnostic accuracy and interobserver agreement on the assessment of OVF versus MVF is moderate at best, irrespective of medical or surgical specialty, years of clinical experience, or hospital type. This result casts doubt on the reliability of using MRI findings together with clinical history as the basis for distinguishing OVF from MVF in routine clinical practice or multicenter studies.

Submitted July 7, 2019; accepted for publication October 7, 2019.

Author contributions: *Study concept:* Arana, Kovacs. *Study design:* Arana, Kovacs, Royuela. *Data collection:* Asenjo, Nagib, Pérez-Aguilera, Dejoz, Cabrera-Zubizarreta, García-Hidalgo, Estremera. *Quality assessment and logistics:* Arana, Dejoz. *Data analysis:* Royuela. *Guidance on clinical practice of vertebral fractures:* Arana. *Writing—original draft and revision:* Arana, Kovacs, Royuela. *Writing—review and editing:* All authors. *Critical revisions:* All authors.

Disclosures: The authors have disclosed that they have not received any financial consideration from any person or organization to support the preparation, analysis, results, or discussion of this article.

Funding: Funding for this work was provided by The Kovacs Foundation.

Correspondence: Estanislao Arana, MD, MHE, PhD, Servicio de Radiología, Fundación Instituto Valenciano de Oncología, C/ Beltrán Báguena, 19, 46009 Valencia, Spain. Email: Estanis.Arana@ext.uv.es

References

1. Kato S, Hozumi T, Yamakawa K, et al. META: an MRI-based scoring system differentiating metastatic from osteoporotic vertebral fractures. Spine J 2015;15:1563–1570.
2. Thawait SK, Kim J, Klufas RA, et al. Comparison of four prediction models to discriminate benign from malignant vertebral compression fractures according to MRI feature analysis. AJR Am J Roentgenol 2013;200:493–502.
3. Wang KC, Jeanmenne A, Weber GM, et al. An online evidence-based decision support system for distinguishing benign from malignant vertebral compression fractures by magnetic resonance imaging feature analysis. J Digit Imaging 2011;24:507–515.
4. Thawait SK, Marcus MA, Morrison WB, et al. Research synthesis: what is the diagnostic performance of magnetic resonance imaging to discriminate benign from malignant vertebral compression fractures? Systematic review and meta-analysis. Spine (Phila Pa 1976) 2012;37:E736–744.
5. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011; 64:96–106.
6. Departamento de Métodos Cuantitativos en Economía y Gestión U de LP de GC. Clasificación de hospitales públicos españoles mediante el uso del análisis cluster [Internet], 2007. Accessed October 10, 2019. Available at: http://www.icmbd.es/docs/resumenClusterHospitales.pdf.
7. Bilsky MH, Laufer I, Fourney DR, et al. Reliability analysis of the epidural spinal cord compression scale. J Neurosurg Spine 2010;13:324–328.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
9. Curtis JR, Taylor AJ, Matthews RS, et al. "Pathologic" fractures: should these be included in epidemiologic studies of osteoporotic fractures? Osteoporos Int 2009;20:1969–1972.
10. Shankar PR, Kaza RK, Al-Hawary MM, et al. Impact of clinical history on maximum PI-RADS version 2 score: a six-reader 120-case sham history retrospective evaluation. Radiology 2018;288:158–163.
11. Levine D, Bankier AA, Halpern EF. Submissions to radiology: our top 10 list of statistical errors. Radiology 2009;253:288–290.
12. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA 2015; 313:1122–1132.
13. Sherbeck JP, Zhao L, Lieberman RW. High variability in lymph node counts among an international cohort of pathologists: questioning the scientific validity of node counts. J Natl Compr Canc Netw 2018;16:395–401.
14. Itri JN, Patel SH. Heuristics and cognitive error in medical imaging. AJR Am J Roentgenol 2018;210:1097–1105.
15. Urrutia J, Besa P, Morales S, et al. Does the META score evaluating osteoporotic and metastatic vertebral fractures have enough agreement to be used by orthopaedic surgeons with different levels of training? Eur Spine J 2018;27:2577–2583.
16. Mauch JT, Carr CM, Cloft H, et al. Review of the imaging features of benign osteoporotic and malignant vertebral compression fractures. AJNR Am J Neuroradiol 2018;39:1584–1592.
17. Sung JK, Jee WH, Jung JY, et al. Differentiation of acute osteoporotic and malignant compression fractures of the spine: use of additive qualitative and quantitative axial diffusion-weighted MR imaging to conventional MR imaging at 3.0 T. Radiology 2014;271:488–498.
18. Besa P, Urrutia J, Campos M, et al. The META score for differentiating metastatic from osteoporotic vertebral fractures: an independent agreement assessment. Spine J 2018;18:2074–2080.
19. Shank CD, Lepard JR, Walters BC, et al. Towards evidence-based guidelines in neurological surgery. Neurosurgery 2019;85:613–621.
20. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA 2004;292:1602–1609.
21. Carney PA, Cook AJ, Miglioretti DL, et al. Use of clinical history affects accuracy of interpretive performance of screening mammography. J Clin Epidemiol 2012;65:219–230.
22. Waite S, Scott J, Gale B, et al. Interpretive error in radiology. AJR Am J Roentgenol 2017;208:739–749.
23. Jarvik JG, Deyo RA. Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology 2009;250:15–17.
24. Wood KB, Khanna G, Vaccaro AR, et al. Assessment of two thoracolumbar fracture classification systems as used by multiple surgeons. J Bone Joint Surg Am 2005;87:1423–1429.
25. Arana E, Kovacs FM, Royuela A, et al. Agreement in the assessment of metastatic spine disease using scoring systems. Radiother Oncol 2015; 115:135–140.
26. Arana E, Kovacs FM, Royuela A, et al. Agreement in metastatic spinal cord compression. J Natl Compr Canc Netw 2016;14:70–76.
27. Fisher CG, Versteeg AL, Schouten R, et al. Reliability of the spinal instability neoplastic scale among radiologists: an assessment of instability secondary to spinal metastases. AJR Am J Roentgenol 2014;203:869–874.
28. Arana E, Kovacs FM, Royuela A, et al. Spine instability neoplastic score: agreement across different medical and surgical specialties. Spine J 2016; 16:591–599.
29. Khan L, Mitera G, Probyn L, et al. Inter-rater reliability between musculoskeletal radiologists and orthopedic surgeons on computed tomography imaging features of spinal metastases. Curr Oncol 2011;18:e282–287.

See JNCCN.org for supplemental online content.

JNCCN

National Comprehensive Cancer Network®

JOURNAL OF THE NATIONAL COMPREHENSIVE CANCER NETWORK

Supplemental online content for:

Metastatic Versus Osteoporotic Vertebral Fractures on MRI: A Blinded, Multicenter, and Multispecialty Observer Agreement Evaluation

Estanislao Arana, MD, MHE, PhD; Francisco M. Kovacs, MD, PhD; Ana Royuela, PhD; Beatriz Asenjo, MD, PhD; Fatima Nagib, MD; Sandra Pérez-Aguilera, MD; María Dejoz, BEng; Alberto Cabrera-Zubizarreta, MD; Yolanda García-Hidalgo, MD, PhD; Ana Estremera, MD, PhD; for the Spanish Back Pain Research Network Task Force for the Improvement of Inter-Disciplinary Management of Spinal Metastasis

J Natl Compr Canc Netw 2020;18(3):267–273

eTable 1. Sequences for MRI Examinations

Pulse Sequence	TR/TE (ms)	FOV (mm)	MAX	NAV	Thickness (mm)	Comments	
Localizer	30/10	400	128 × 128	1	10	Flip angle 50°	Gradient echo
Sagittal T1	440–550/14–20	270	156–307 × 192–512	2	4	1.3–0.4 mm gap	Spin-echo
Sagittal T2	3,300–2,896/102.9–120	270	156–307 × 192–512	2	4	1.3–0.4 mm gap	Turbo spin-echo imaging, 12-echo train length
Sagittal STIR	3,000/45/150 (inversion time)	270	156–307 × 192–512	2	4–6	1.3–0.4 mm gap	Turbo spin-echo imaging, 12-echo train length
Axial T2	3,040–2,896/103–120	180	224–190 × 256–512	3	4	0.4 mm gap	Turbo spin-echo imaging, 5-echo train length

Abbreviations: FOV, field of view; MAX, matrix; NAV, number of signals acquired; STIR, short inversion time inversion-recovery; TE, echo time; TR, repetition time.

eTable 2. Imaging Findings Assessed

Imaging Finding	Possible Values
Pattern of signal abnormalities (pattern of replacement of normal vertebral signal with bone marrow edema)	"Partially or completely" vs "showing a bandlike pattern"
Horizontal fracture line on fluid-sensitive sequence (STIR) or T2-weighted images	"Yes" vs "no"
Deposit-like appearance of pedicle involvement	"Yes" vs "no"
Convexity of posterior vertebral body border (bulging posterior cortex)	"Yes" vs "no"
Posterosuperior retropulsion	"Yes" vs "no"
Symmetry of signal intensity changes	"Symmetrical" vs "asymmetrical"

Abbreviation: STIR, short inversion time inversion-recovery.

eTable 3. Interobserver Agreement on Imaging Findings

	Kappa (95% CI)
Radiology (n=9)	
Pattern of signal abnormalities	0.410 (0.351–0.473)
Horizontal fracture line	0.352 (0.277–0.432)
Deposit-like appearance of pedicle involvement	0.476 (0.422–0.534)
Bulging posterior cortex	0.602 (0.545–0.661)
Posterosuperior retropulsion	0.367 (0.312–0.424)
Symmetry of signal intensity changes	0.277 (0.229–0.327)
Neurosurgery (n=7)	
Pattern of signal abnormalities	0.428 (0.365–0.495)
Horizontal fracture line	0.130 (0.087–0.176)
Deposit-like appearance of pedicle involvement	0.473 (0.409–0.539)
Bulging posterior cortex	0.400 (0.339–0.464)
Posterosuperior retropulsion	0.445 (0.390–0.502)
Symmetry of signal intensity changes	0.267 (0.213–0.324)
Orthopedic surgery (n=5)	
Pattern of signal abnormalities	0.327 (0.270–0.386)
Horizontal fracture line	0.198 (0.145–0.253)
Deposit-like appearance of pedicle involvement	0.412 (0.340–0.487)
Bulging posterior cortex	0.104 (0.064–0.144)
Posterosuperior retropulsion	0.533 (0.467–0.602)
Symmetry of signal intensity changes	0.163 (0.109–0.219)
Radiation oncology (n=4)	
Pattern of signal abnormalities	0.355 (0.280–0.433)
Horizontal fracture line	0.326 (0.242–0.412)
Deposit-like appearance of pedicle involvement	0.416 (0.341–0.493)
Bulging posterior cortex	0.635 (0.561–0.711)
Posterosuperior retropulsion	0.101 (0.047–0.155)
Symmetry of signal intensity changes	0.388 (0.314–0.465)

eTable 4. Intraobserver Agreement on Imaging Findings

	Median Kappa (IQR)
Radiology (n=9)	
Pattern of signal abnormalities	0.722 (0.606–0.764)
Horizontal fracture line	0.639 (0.472–0.721)
Deposit-like appearance of pedicle involvement	0.707 (0.624–0.732)
Bulging posterior cortex	0.768 (0.640–0.800)
Posterosuperior retropulsion	0.673 (0.624–0.731)
Symmetry of signal intensity changes	0.575 (0.383–0.646)
Neurosurgery (n=7)	
Pattern of signal abnormalities	0.754 (0.533–0.894)
Horizontal fracture line	0.657 (0.458–0.914)
Deposit-like appearance of pedicle involvement	0.653 (0.527–0.914)
Bulging posterior cortex	0.844 (0.495–0.969)
Posterosuperior retropulsion	0.689 (0.617–0.941)
Symmetry of signal intensity changes	0.597 (0.402–0.902)
Orthopedic surgery (n=5)	
Pattern of signal abnormalities	0.549 (0.510–0.555)
Horizontal fracture line	0.457 (0.399–0.515)
Deposit-like appearance of pedicle involvement	0.504 (0.460–0.549)
Bulging posterior cortex	0.682 (0.618–0.693)
Posterosuperior retropulsion	0.712 (0.587–0.719)
Symmetry of signal intensity changes	0.409 (0.360–0.460)
Radiation oncology (n=4)	
Pattern of signal abnormalities	0.646 (0.603–0.808)
Horizontal fracture line	0.486 (0.433– 0.712)
Deposit-like appearance of pedicle involvement	0.666 (0.592–0.797)
Bulging posterior cortex	0.724 (0.639–0.846)
Posterosuperior retropulsion	0.586 (0.429–0.775)
Symmetry of signal intensity changes	0.584 (0.517–0.753)

Abbreviation: IQR, interquartile range.

eTable 5. Diagnostic Accuracy[a]

	N	Median Kappa (IQR)
All readers		
Cancer history undisclosed	25	0.437 (0.326–0.511)
Cancer history disclosed	25	0.443 (0.398–0.526)
Specialty		
Neurosurgery		
Cancer history undisclosed	7	0.327 (0.230–0.511)
Cancer history disclosed	7	0.411 (0.314–0.534)
Radiation oncology		
Cancer history undisclosed	4	0.446 (0.348–0.507)
Cancer history disclosed	4	0.435 (0.354–0.490)
Orthopedic surgery		
Cancer history undisclosed	5	0.368 (0.325–0.445)
Cancer history disclosed	5	0.398 (0.311–0.444)
Radiology		
Cancer history undisclosed	9	0.437 (0.414–0.525)
Cancer history disclosed	9	0.484 (0.443–0.526)
Hospital category (complexity)[b]		
Category 2		
Cancer history undisclosed	2	0.381 (0.325–0.437)
Cancer history disclosed	2	0.372 (0.311–0.433)
Category 3		
Cancer history undisclosed	9	0.470 (0.403–0.525)
Cancer history disclosed	9	0.484 (0.410–0.534)
Category 4		
Cancer history undisclosed	7	0.445 (0.327–0.565)
Cancer history disclosed	7	0.437 (0.411–0.543)
Category 5		
Cancer history undisclosed	7	0.413 (0.281–0.426)
Cancer history disclosed	7	0.443 (0.359–0.526)
Years of experience		
≤7		
Cancer history undisclosed	7	0.403 (0.325–0.437)
Cancer history disclosed	7	0.411 (0.359–0.491)
8–13		
Cancer history undisclosed	6	0.397 (0.253–0.445)
Cancer history disclosed	6	0.421 (0.314–0.526)
≥14		
Cancer history undisclosed	12	0.491 (0.428–0.543)
Cancer history disclosed	12	0.477 (0.435–0.554)

Abbreviations: IQR, interquartile range; MVF, metastatic vertebral fracture; OVF, osteoporotic vertebral fracture.

[a] Defined as the concordance between each reader's diagnosis at the first round (OVF vs MVF) and the reference diagnosis (established through biopsy or follow-up >6 months).

[b] Based on size, availability of high-tech medical equipment and procedures, and degree of educational activity. No readers from category 1 hospitals (simplest) were included in this study.

eTable 6. Diagnostic Accuracy[a] Depending on Presence of Preexisting Fractures and Disclosure of Clinical History

	Median Kappa (IQR)
Cases without preexisting fractures	
Before clinical history of cancer was disclosed	0.452 (0.387–0.509)
After clinical history of cancer was disclosed	0.462 (0.407–0.570)
Cases with preexisting fractures	
Before clinical history of cancer was disclosed	0.286 (0.183–0.396)
After clinical history of cancer was disclosed	0.331 (0.219–0.368)

Abbreviations: IQR, interquartile range; MVF, metastatic vertebral fracture; OVF, osteoporotic vertebral fracture.

[a] Diagnostic accuracy is defined as the concordance between each reader's diagnosis at the first round (OVF vs MVF) and the reference diagnosis (established through biopsy or follow-up >6 months).

eAppendix 1.

Members of the Spanish Back Pain Research Network Task Force for the Improvement of Inter-Disciplinary Management of Spinal Metastasis (in alphabetical order)

Ana Alonso[1,2]; Marco Antonio Álvarez[1,3]; Luis Álvarez-Galovich[1,4]; Aida Antuña[1,3]; Joaquín Cabrera[1,5]; Carlos Casillas[1,6]; Gregorio Catalán[7,8]; Diego Dualde[7,9]; Nicomedes Fernández-Baillo[7,10]; Antonio Ferreiro[7,11]; Pilar Ferrer[1,12]; Sara García-Duque[7,13]; Cristina García-Villar[7,14]; Ovidio Hernando-Requejo[1,15]; Laín Ibáñez[1,16]; Ana Lersundi[1,17]; Marta Manero[1,18]; Antonio Martín[1,19]; Julio César Palomino[7,20]; Luis A. Pérez-Romasanta[1,21]; Julio Plata-Bello[1,22]; Raquel Prada[1,20]; Héctor Roldán[1,22]; Luis Maria Romero-Muñoz[1,23]; Félix Tomé-Bermejo[1,4]; Vicente Vanaclocha[1,24]; and Joaquín Zamarro[7,25]

[1]Spanish Back Pain Research Network, Kovacs Foundation, Palma de Mallorca, Spain
[2]Hospital Universitario Rey Juan Carlos, Móstoles, Madrid, Spain
[3]Hospital Universitario Central de Asturias, Asturias, Spain
[4]Hospital Universitario Fundación Jiménez Díaz, Madrid, Spain
[5]Hospital Universitario de Badajoz, Badajoz, Spain
[6]Hospital Jaume I, Castellón, Spain
[7]Unidad de la Espalda Kovacs, Hospital Universitario HLA-Moncloa, Madrid
[8]Hospital de Cruces, Baracaldo, Spain
[9]Hospital Clínico Universitario de Valencia, Valencia, Spain
[10]Hospital La Paz, Madrid, Spain
[11]Hospital de Madrid, HM Hospitales, Madrid, Spain
[12]Hospital Intermutual de Levante, San Antonio de Benagéber, Valencia, Spain
[13]Hospital Universitario HM Sanchinarro, Madrid, Spain
[14]Hospital Universitario Puerta del Mar, Cádiz, Spain
[15]Hospital Universitario HM Puerta del Sur, Móstoles, Madrid, Spain
[16]Hospital Universitario 12 de Octubre, Madrid, Spain
[17]Hospital Universitario Donostia, Donostia, Gipuzkoa, Spain
[18]Clínica Vistahermosa, Alicante, Spain
[19]Hospital Doctor Peset, Valencia, Spain
[20]Hospital POVISA, Vigo, Spain
[21]Hospital Universitario de Salamanca, Salamanca, Spain
[22]Hospital Universitario de Canarias, Santa Cruz de Tenerife, Spain
[23]Hospital Nacional de Parapléjicos, Toledo, Spain
[24]Hospital General Universitario de Valencia, Valencia, Spain
[25]Hospital Universitario Virgen de la Arrixaca, Murcia, Spain