

# Inter- and intra-observer variability associated with the use of the Mirels' scoring system for metastatic bone lesions

Ruairi F. Mac Niocaill · John F. Quinlan ·  
Robert D. Stapleton · Brian Hurson · Sean Dudeney ·  
Gary C. O'Toole

Received: 16 November 2009 / Revised: 14 December 2009 / Accepted: 15 December 2009 / Published online: 19 January 2010  
© Springer-Verlag 2010

**Abstract** Metastatic bone disease is increasing in association with ever-improving medical management of osteophylic malignant conditions. The precise timing of surgical intervention for secondary lesions in long bones can be difficult to determine. This paper aims to evaluate a classic scoring system. All radiographs were examined twice by three orthopaedic oncologists and scored according to the Mirels' scoring system. The Kappa statistic was used for the purpose of statistical analysis. The results show agreement between observers ( $\kappa=0.35$ – $0.61$ ) for overall scores at the two time intervals. Inter-observer agreement was also seen with subset analysis of size ( $\kappa=0.27$ – $0.60$ ), site ( $\kappa=0.77$ – $1.0$ ) and nature of the lesion ( $\kappa=0.55$ – $0.81$ ). Similarly, low levels of intra-observer variability were noted for each of the three surgeons ( $\kappa=0.34$ ,  $0.39$ , and  $0.78$ , respectively). These results indicate a reliable, repeatable assessment of bony metastases. We continue to advocate its use in the management of patients with long bone metastases.

## Introduction

The skeleton is the most common organ to be affected by metastatic cancer with a predilection of the common cancers to metastasise to bone [4]. Tumour registry figures suggest that incidence of bone metastases is increasing, with breast being the most common causative histology and the femur and spine the most common sites [20]; in addition, bone metastases have been found to be the first sign of disease recurrence in a small number of patients [15]. An estimated 350,000 people die with bone metastases in the United States each year [14]. The management of metastatic deposits in long bones has long been a source of discussion. Many authors have proposed methods with which to identify those lesions at risk of causing pathological fractures based on radiological and clinical factors [2, 9–11, 16, 17, 19]. The basis of these methods of prediction generally take into account the size of the lesion, whether it involves a weight-bearing bone and whether the lesion is lytic or sclerotic in nature.

The most widely accepted of these predictive systems is that of Mirels [13], who proposed a scoring system based on pain intensity, site, type (lytic, mixed or blastic) and amount of bony involvement (Table 1). Mirels' system is widely used. It is validated in the original study using a small sample size (38 patients) and has been subject to independent validation in only one other significant review [5]. This review by Damron et al. is itself limited by relatively small sample size ( $n=12$ ) and the use of simplified clinical histories requiring physician assessment pain severity based on written information provided.

The inclusion of physician rated pain severity in clinical scoring systems is problematic as pain is a subjective experience with both physical and psychosocial elements that

---

R. F. Mac Niocaill · J. F. Quinlan · B. Hurson · S. Dudeney ·  
G. C. O'Toole  
Department of Orthopaedic Surgery,  
St. Vincent's University Hospital,  
Elm Park,  
Dublin 4, Republic of Ireland

R. D. Stapleton  
Department of Mathematics, National University of Ireland,  
Maynooth,  
Co. Kildare, Republic of Ireland

R. F. Mac Niocaill (✉)  
15 Hillsborough, Laraghcon, Lucan,  
Co. Dublin, Republic of Ireland  
e-mail: ruairi99@hotmail.com

**Table 1** Mirels' scoring system

Parameter	Score		
	1	2	3
Site	Upper limb	Lower limb	Peritrochanter
Pain	Mild	Moderate	Severe
Lesion	Blastic	Mixed	Lytic
Size	<1/3	1/3–2/3	>2/3

are difficult to quantify objectively. Furthermore, the paucity of empirical data using validated pain assessments for bone pain also complicates the matters [6]. While the importance of pain severity in the assessment of fracture risk is generally accepted, it is however not absolute as two significant studies have shown [8, 12]. Keene et al. [12], whose paper is one of the largest on the subject, found that pain was not a significant predictor of fracture. Damron et al. [5] also showed in their intra- and inter-observer concordance study that pain was the factor which showed greatest variability.

The aim of this study was to independently evaluate the Mirels' scoring system as applied to a cohort of bony metastatic disease in terms of inter- and intra-observer variability with the objective of obtaining data relating to its suitability for application as an 'off the shelf' aid to decision making in orthopaedic oncology. It is a basic premise of predictive scoring systems that they show satisfactory intra- and inter-observer reliability from both a clinical and academic point of view. In order for treatment decisions to be logical and consistent both within and between treating institutions and in order for reported treatment results to be valid, it is vital to have a predictive tool that produces similar results between individual clinicians and with repeated use. To remove the potential for bias caused by patient or physician rated pain severity, only the radiological features of the system were evaluated, thereby giving a real sense of the reproducibility of this system using only its most objective elements.

## Materials and methods

### Patients

Surgical, oncology and HIPE (hospital in-patient enquiry) records from the period between January 2005 and June 2007 inclusive were examined in an effort to identify patients with long bone metastases, and a retrospective chart and radiological review was carried out.

Criteria for selection and inclusion in the study were:

1. A known histologically proven primary neoplasm
2. A synchronous metastatic lesion present in a long bone, diagnosed radiologically

3. No fracture or history of fracture through this lesion
4. A comprehensive series of pre-fracture, pre-intervention radiographs were selected

Patients who had undergone adjuvant therapy were excluded as were those in which no histologically proven primary were identified.

Radiographs showing 35 lesions in 28 patients who met the selection criteria were retrieved. A patient database containing data regarding age, gender, histology and sites affected was created. Only those with pre-treatment images were selected, in particular no post radiotherapy images were used.

The radiographs were reviewed by three fellowship trained orthopaedic surgical oncologists (BH, SD & GOT) using a standard proforma assessment sheet containing the Mirels' scoring system table. No clinical data were provided and the reviewers rated the radiological features only. This review process was repeated three weeks later using the same radiographs with altered sequence and labelling. The surgeons were blinded to patient identity and no patients currently being treated in the unit were included. Scores were recorded out of a maximum of nine rather than 12 as pain was not considered in this study.

The mean age (mean  $\pm$  standard deviation) of the patients in this study was  $62.3 \pm 11.1$  years (range 39–81 years). There were 11 male and 17 female patients. The bones affected by metastases were the femur ( $n=26$ ), humerus ( $n=6$ ) and tibia ( $n=3$ ). The primary neoplasms represented in the study cohort were: breast carcinoma ( $n=11$ ), small cell lung carcinoma ( $n=6$ ), multiple myeloma ( $n=5$ ), prostate carcinoma ( $n=4$ ), non-small cell lung carcinoma ( $n=3$ ), renal cell carcinoma ( $n=3$ ), thyroid carcinoma ( $n=2$ ), colorectal carcinoma ( $n=1$ ) and alveolar soft part sarcoma ( $n=1$ ).

### Statistical analysis

The data were analysed for both inter- and intra-observer agreement. For inter-observer agreement, the initial overall score and scores for site, size and nature of lesion were compared across each pair of surgeons. As such, scores for surgeon 1 were compared with scores for surgeon 2 and similarly comparisons for surgeons 1 and 3 with surgeons 2 and 3. The scores assigned for the second observational time-point were similarly compared with each comparison performed using the Kappa statistic. The Kappa statistic considers the null hypothesis of no agreement versus the alternative hypothesis of agreement beyond what would be expected by chance, with a Kappa statistic of 0 indicating agreement that could be expected by chance and a Kappa statistic of 1 indicating complete agreement.

For intra-observer variability, the initial overall score and scores for site, size and nature of lesion were compared to the

second recorded overall score and score for site, size and nature of lesion, respectively, again using the Kappa statistic.

A *p*-value of less than 0.05 was considered to be statistically significant. All statistical analyses were conducted using the statistical package SPSS 14.0 (SPSS Inc., Chicago, Ill, USA).

## Results

### Results for inter-observer analysis

All results were reported at a significance level of  $p < 0.001$  except where specifically stated.

For the overall score comparisons, there was evidence of agreement beyond that expected by chance when comparing surgeons 1 and 2 at both time points ( $\kappa=0.350$  and  $0.505$ , respectively) and surgeons 2 and 3 at both time points ( $\kappa=0.404$  and  $0.610$ , respectively). Surgeons 1 and 3 only demonstrated significant agreement at the second time point ( $\kappa=0.485$ ).

In relation to site score for the first scoring of the X-rays (first observational time-point), there was significant agreement when comparing surgeons 1 and 2 ( $\kappa=0.818$ ), surgeons 1 and 3 ( $\kappa=0.770$ ) and surgeons 2 and 3 ( $\kappa=0.955$ ). Similar results were found at the second observational time-point with concurrence between scores for surgeons 1 and 2 ( $\kappa=0.863$ ), surgeons 1 and 3 ( $\kappa=0.863$ ) and surgeons 2 and 3 ( $\kappa=1.000$ ).

There was agreement between surgeons 1 and 2 ( $\kappa=0.475$ ), surgeons 1 and 3 ( $\kappa=0.267$ ,  $p=0.024$ ) and surgeons 2 and 3 ( $\kappa=0.521$ ) at the first viewing. A similar pattern was seen at the second observational time-point with a similarity of results when comparing surgeons 1 and 2 ( $\kappa=0.506$ ), surgeons 1 and 3 ( $\kappa=0.596$ ) and surgeons 2 and 3 ( $\kappa=0.558$ ).

For nature of lesion analysis at the first X-ray scoring, there was significant concordance between all observers (surgeons 1 and 2 [ $\kappa=0.814$ ], surgeons 1 and 3 [ $\kappa=0.589$ ] and surgeons 2 and 3 [ $\kappa=0.695$ ]). Similarly, at the second observational time-point, there was again evidence of agreement comparing surgeons 1 and 2 ( $\kappa=0.669$ ), surgeons 1 and 3 ( $\kappa=0.550$ ) and surgeons 2 and 3 ( $\kappa=0.626$ ).

### Results for intra-observer analysis

For surgeon 1, there was evidence of agreement when comparing the first observational and second observational time-points for overall scores ( $\kappa=0.340$ ) as well as comparing the scores for site ( $\kappa=0.765$ ), size ( $\kappa=0.481$ ) and nature of lesion scores ( $\kappa=0.757$ ).

In the case of the second surgeon, there was evidence of agreement when comparing the first observational and second observational time-points for overall scores ( $\kappa=0.392$ ). There

was also similarity for the site ( $\kappa=1.000$ ), size ( $\kappa=0.438$ ) and nature of lesion scores ( $\kappa=0.656$ ).

The observations of surgeon 3 showed agreement when comparing the first observational and second observational time-points for overall score ( $\kappa=0.788$ ), site ( $\kappa=0.955$ ), size ( $\kappa=0.561$ ) and nature of lesion scores ( $\kappa=0.766$ ).

The results for intra-observer analysis are shown in Table 2.

## Discussion

Mirels, in his paper of 1989 [13], presented a proposed scoring system to quantify the risk of sustaining a pathological fracture through a metastatic lesion in a long bone. He did this by performing a retrospective analysis of 78 lesions in 38 patients that had been irradiated without prophylactic fixation. The ensuing scoring system had a maximum score of 12 that could be attained with individual scores of up to 3 for the four subgroups of site, pain, lesion size and whether the lesion was lytic, mixed or sclerotic. The conclusion of his work suggested that long bones with lesions that score 9 or more should undergo prophylactic fixation.

Patients undergoing fixation of pathological fractures benefit from these procedures in terms of mobility and reduction in local pain [18]. Prevention of fracture by prophylactic fixation offers both technical and patient related benefits. In terms of operative procedures, a prophylactic fixation is considered to be of a lesser magnitude than having to fix an established pathological fracture [1, 3, 8, 21]. Furthermore, in relation to the patient, prophylactic fixation has been associated with pain relief with resultant improvement in the quality of life and restoration of ambulation [19] as well as a low complication rate [7].

As previously discussed, the reliability of pain as a predictor of fracture has been questioned and may indeed act as a confounding factor in the prediction of impending

**Table 2** Intra-observer analysis for all surgeons

Surgeon	Subset	$\kappa$ -statistic	<i>p</i> -value
Surgeon 1	Overall	0.340	<0.001
	Site	0.765	<0.001
	Size	0.481	<0.001
	Nature	0.757	<0.001
Surgeon 2	Overall	0.392	<0.001
	Site	1.0	<0.001
	Size	0.438	<0.001
	Nature	0.656	<0.001
Surgeon 3	Overall	0.788	<0.001
	Site	0.955	<0.001
	Size	0.561	<0.001
	Nature	0.766	<0.001

pathological fractures through metastases [8, 12] by significantly altering the scores recorded. As such, this study concentrated exclusively on the radiological components of the Mirels' system, thereby assessing only the most objective elements of the scoring system. We wished to evaluate the intra- and inter-observer reliability of the system as appropriate levels of both are highly desirable in any clinical scoring system and are in reality a prerequisite for acceptability of any clinical test. This is true in particular when reporting results of treatment in the scientific literature in which the validity of results and conclusions rely on such "like for like" comparisons.

The importance of both site of the lesion and its association with pain generation as well as better understanding of fracture risk in bone appearing sclerotic is acknowledged but is beyond the scope of this paper.

In this study we have also facilitated the application of the scoring system to a relatively broad array of pathology in terms of histology, site and type of lesion than has been the case in other assessments of the Mirels' scoring system to date. In doing so, we hope to have provided an improved understanding of the reliability and reproducibility possible with the use of this scoring system.

The results have shown that when applied by experienced orthopaedic surgical oncologists, there is statistically significant inter- and intra-observer agreement across the spectrum of disease patterns. Analysis of subgroups in relation to time-points, size, site, nature of lesions and high or low scoring patterns similarly recorded a high level of agreement throughout the study. These results compare favourably to the only other significant independent appraisal of the reliability of the Mirels system, which was made by Damron et al. [5].

This paper excludes pain in the assessment. This approach is potentially controversial as pain is integral to many of the scoring systems used in this area. Our objective was, however, to identify the reproducibility of the radiological features of the Mirels score when applied by experienced clinicians, as no empirical data relating to this vital element exists in the scientific literature to date. We acknowledge the potential for bias caused by the relatively short re-review interval; however, we feel that, overall, adequate precautions to minimise this factor were taken.

In conclusion, the results of this study would advocate the application of the radiological components of the Mirels' scoring system as reliable and repeatable as applied to this cohort of patients. While the pitfalls of the pain subset in altering the score are documented and recognised, this paper would support the continued and regular use of the Mirels scoring system in the management of patients with malignant bone disease.

## References

1. Bonargio BC, Rubin P (1967) Non-union of pathological fracture after irradiation therapy. *Radiology* 88:889–898
2. Bremner RA, Jelliffe AM (1958) The management of pathological fracture of the major long bones from metastatic cancer. *J Bone Jt Surg* 40B:652–659
3. Bunting RW, Boublik M, Blevins FT, Dame CC, Ford LA, Lavine LS (1992) Functional outcome of pathological fracture secondary to malignant disease in a rehabilitation hospital. *Cancer* 69:98–102
4. Coleman RE (1997) Skeletal complications of malignancy. *Cancer* 80:1588–1594
5. Damron TA, Morgan H, Prakash D, Grant W, Aronowitz J, Heiner J (2003) Critical evaluation of Mirels rating system for impending pathological fractures. *Clin Orthop* 415S:201–207
6. Dawson R, Currow D, Stevens G, Morgan G, Barton M (1999) Radiotherapy for bone metastases: a critical appraisal of outcome measures. *J Pain Symptom Manage* 17:208–218
7. Dijkstra S, Wiggers T, van Geel BN, Boxma H (1994) Impending and actual pathological fractures in patients with bone metastases of the long bones. A retrospective study of 233 surgically treated fractures. *Eur J Surg* 160:535–542
8. Fidler M (1973) Prophylactic internal fixation of secondary neoplastic deposits in long bones. *Br Med J* 1:341–343
9. Habermann ET, Sachs R, Stern RE, Hirsh DM, Anderson WJ (1982) The pathology and treatment of metastatic disease of the femur. *Clin Orthop Relat Res* 169:70–82
10. Harrington KD (1982) New trends in management of lower extremity metastases. *Clin Orthop* 169:53–61
11. Harrington KD, Sim FH, Enis JE, Johnson JD, Dick HM, Gristina AG (1976) Methylmethacrylate as an adjunct in internal fixation of pathological fractures. *J Bone Jt Surg* 58A:1047–1055
12. Keene JS, Sellinger MD, McBeath AA, Engber WD (1986) Metastatic breast cancer in the femur: a search for the lesion at risk of fracture. *Clin Orthop* 203:282–288
13. Mirels H (1989) Metastatic disease in long bones: a proposed scoring system for diagnosing impending pathological fractures. *Clin Orthop* 249:256–264
14. Mundy GR (2002) Metastasis to bone: causes, consequences and therapeutic opportunities. *Nat Rev Cancer* 2:584–593
15. Papadakis SA, Mitsitsikas TC, Markakidis S, Minas MK, Tripsiannis G, Tentes AA (2004) The development of bone metastases as the first sign of metastatic spread in patients with primary solid tumours. *Int Orthop* 28(2):102–105
16. Parrish FF, Murray JA (1974) Surgical management of secondary neoplastic fractures about the hip. *Orthop Clin North Am* 5(4):887–901
17. Perez CA, Bradfield JS, Morgan HC (1972) Management of pathological fractures. *Cancer* 29:684–693
18. Sarahrudi K, Hora K, Heinz MS, Vecsei V (2006) Treatment results of pathological fractures of the long bones: a retrospective analysis of 88 patients. *Int Orthop* 30(6):519–524
19. Schurman DJ, Amstutz HC (1973) Orthopaedic management of patients with metastatic carcinoma of the breast. *Surg Gynecol Obstet* 137:831–836
20. Toma CD, Dominkus M, Nedelcu T, Abdolvahab F, Assadian O, Krepler P, Kotz R (2007) Metastatic bone disease: a 36-year single centre trend-analysis of patients admitted to a tertiary orthopaedic surgical department. *J Surg Oncol* 96(5):404–410
21. Ward WG, Spang J, Howe D, Gordan S (2000) Femoral recon nails for metastatic disease: indications, technique and results. *Am J Orthop* 29(9 Suppl):34–42