Classifications for Radiographic Evaluation of Lytic Bone Lesions have Poor Inter- and Intra-observer Agreement

Taylor Willenbring, Sarah Papa, Salvatore Cavallaro, Kenneth Mann, Timothy Damron
SUNY Upstate Medical University, Syracuse NY
DamronT@upstate.edu

Disclosures: Taylor Willenbring (N), Sarah Papa (N), Salvatore Cavallaro (N), Kenneth Mann (N), Timothy Damron (N)

INTRODUCTION: Lytic lesions of bone are encountered within all orthopedic specialties, and therefore concise methods of describing them are essential for effective communication between practitioners and to inform the decision regarding need for biopsy. We studied three classification systems of lytic lesions on radiographs: (1) the original classification by Lodwick[1], (2) a modified Lodwick classification[2], and (3) a simplified, functional classification for benign tumors by Enneking[3]. This study evaluates the inter-observer reliability and intra-observer reproducibility of these three classification systems as used by orthopedic trainees and physicians. We hypothesized that intra-observer reproducibility would be good but overall inter-observer reliability would be poor, improving with training level and that the simplified classification would have the best reproducibility.

METHODS: Forty-eight case sets of deidentified radiographs of lytic osseous lesions with known diagnoses were selected from an orthopedic oncology practice to represent a broad array of tumors and radiographic appearances. Each set included at least two orthogonal views of the lesion from initial presentation without prior biopsy or oncologic treatment. Lesions were then classified according to the Lodwick classification[1], the modified Lodwick classification[2], and a functional classification[3] twice by each participant with a minimum two-week gap between sessions. All participants were provided a reference sheet with classification details for use throughout. Twenty participants (one third-year medical student, 18 residents, one orthopedic oncologist) completed the two sessions. Inter-observer reliability was calculated using both Fleiss' kappa and Krippendorff's alpha, treating the classifications as nominal and ordinal rankings, respectively. Intra-observer reproducibility was also calculated using Fleiss' kappa and Krippendorff's alpha. Linear regression models were then used to determine the effect of training level on ability to apply the classifications consistently.

RESULTS: Inter-observer reliability was poor for all three classifications as demonstrated by agreement of 39% (κ 0.23), 39% (κ 0.25), and 53% (κ 0.28) for the Lodwick, modified Lodwick, and functional classifications, respectively (Table 1). None of the three achieved "moderate agreement" levels for reliability according to kappa interpretations by either Landis[4] or McHugh[5]. When the classifications were treated as ordinal data, there was not an improvement in the overall agreement (α 0.54, 0.48, 0.45, respectively). Traditionally, data is considered reliable when Krippendorff's $\alpha > 0.8$ with higher stringency required for higher risk scenarios and discarding those with $\alpha < 0.67$[6]. Intra-observer reproducibility also lacked strong agreement, although kappa values were improved (κ 0.42-0.45) relative to the inter-observer measures (Table 1). Training level had no effect on the ability to reproducibly classify lesions using any of the three classifications ($R^2 < 0.2$, $p>0.05$ for all classifications, Figure 1). Comparison of classification systems with respect to intra-observer reproducibility showed Krippendorff's alpha for the Lodwick (α 0.72), modified Lodwick (α 0.69), and functional classification (α 0.63). Self-agreement for individuals was quite variable, ranging from 39-78%.

DISCUSSION: One of the primary utilities of a classification is its ability to allow effective communication between physicians. Our data demonstrate that three commonly used classifications for osseous lytic lesions on radiographs are not reliable nor reproducible. It was noted that the consistency of classification was highly variable depending on the individual lesions' characteristics. These classifications may be useful for certain lesions but are unable to be applied broadly across a wide array of lesion types. Interestingly, there was no association between orthopedic experience and intra-observer reproducibility, supporting the notion that the descriptions themselves are poorly applied to some lesions. These deficiencies may be improved with AI applications in the future.

SIGNIFICANCE: This study demonstrates the poor inter-observer reliability and intra-observer reproducibility of three classifications used to radiographically describe lytic lesions of bone. As they currently stand, there is no reliable method to classify these lesions and communicate about them effectively. Some lesions are more easily described than others.

REFERENCES:

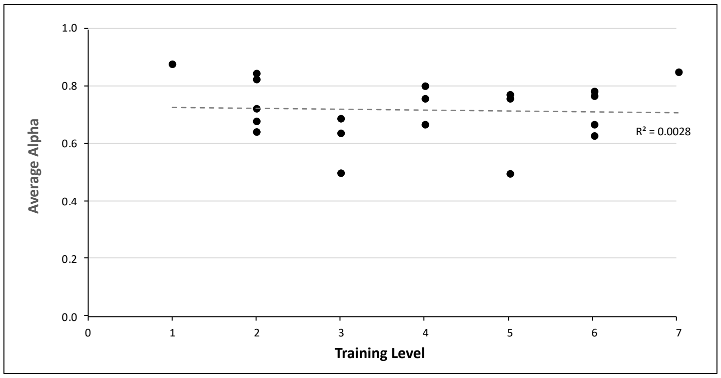
1. Lodwick et al., 1980. 2. Costelloe et al. 2012, 3. Enneking et al. 1980, 4. Landis et al. 1977, 5. McHugh et al. 2012. 6. Krippendorff 2004.

IMAGES AND TABLES:

Table 1. Agreement kappa and alpha values for each classification.

Classification	Inter-rater Reliability		Intra-rater Reproducibility	
	Kappa	*Alpha*	*Kappa*	*Alpha*
Lodwick	0.23	0.54	0.45	0.71
Modified Lodwick	0.25	0.48	0.44	0.69
Functional	0.28	0.45	0.43	0.63

Figure 1. Overall reproducibility using Krippendorff's alpha, categorized by training level (1=Student, 2=PGY-1 … 6=PGY-5, 7=attending).


Average Alpha
R² = 0.0028
Training Level