SPINE Volume 34, Number 4, pp 384–391

An Assessment of the Reliability of the Enneking and Weinstein-Boriani-Biagini Classifications for Staging of Primary Spinal Tumors by the Spine Oncology Study Group

Patrick Chan, MD,* Stefano Boriani, MD,† Daryl R. Fourney, MD,‡ Roberto Biagini, MD,§ Mark B. Dekutoski, MD,¶ Michael G. Fehlings, MD, PhD,‖ Timothy C. Ryken, MD,** Ziya L. Gokaslan, MD,†† Frank D. Vrionis, MD, PhD, MPH,‡‡ James S. Harrop, MD,§§ Meic H. Schmidt, MD,¶¶ Luis R. Vialle, MD,‖‖ Peter C. Gerszten, MD,*** Laurence D. Rhines, MD,††† Stephen L. Ondra, MD,‡‡‡ Stuart R. Pratt, MS,§§§ and Charles G. Fisher, MD, MHSc*

Study Design. Reliability analysis based on expert panel case series review and grading per the Enneking and Weinstein-Boriani-Biagini classification systems.

Objective. To assess the reliability of the Enneking and Weinstein-Boriani-Biagini classification systems.

Summary of Background Data. The Enneking and Weinstein-Boriani-Biagini (WBB) classifications were developed to stage and facilitate treatment planning in patients with primary spine tumors. To date, their interobserver and intraobserver reliability has not been assessed–a fundamental step in facilitating broader clinical and research use.

Methods. Clinical information, imaging studies, and biopsy results were compiled from 15 selected patients with primary spinal tumors. Eighteen spine surgeons independently estimated and scored the cases for Enneking grade, tumor and metastasis categories, Enneking stage, Enneking-recommended surgical margin, WBB zones and layers, and WBB-recommended surgical procedures, with a second assessment performed after random resorting of cases. Interobserver and intraobserver reliability of each category were assessed by percent agreement or proportional overlap. The Fleiss, Cohen, and Mezzich κ statistics (κ) were then applied, determined by the type of variable analyzed.

Results. The κ statistics for interobserver reliability were 0.82, 0.22, 0.00, 0.57, 0.47, 0.31, 0.58, and 0.54 for the fields of Enneking grade, tumor and metastasis categories, Enneking stage, Enneking-recommended surgical margin, WBB zones and layers, and WBB-recommended surgical procedures, respectively. The κ statistics for intraobserver reliability were 0.97, 0.53, 0.47, 0.82, 0.67, 0.63, 0.79, and 0.79 for the same respective fields. According to Landis and Koch, the ranges of κ values of 0.00 to 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80, and >0.80 imply slight, fair, moderate, substantial, and near-perfect agreement, respectively.

Conclusion. Results indicate moderate interobserver reliability and substantial and near-perfect intraobserver reliability for both the Enneking and WBB classification in terms of staging and guidance for treatment, despite a less than moderate interobserver reliability in interpreting the Enneking local tumor extension and WBB sector. Before incorporating the classifications in the clinical practice and research studies, further work is required to investigate the validity of the classifications.

Key words: intraobserver reliability, primary tumors of spine, classification system, oncological staging. **Spine 2009;34:384–391**

From the * Vancouver General Hospital, Vancouver, Canada; †Ospedale Maggiore, Bologna, Italy; ‡University of Saskatchewan Saskatchewan, Canada; §Instituto Regina Elena, Rome, Italy; ¶Mayo Clinic, Rochester, MN; ‖University of Toronto, Toronto, Canada; **University of Iowa, Iowa City, IA; ††Johns Hopkins University, Baltimore, MD; ‡‡H. Lee Moffitt Cancer Center, Tampa, FL; §§Thomas Jefferson University Hospital, Philadelphia, PA; ¶¶University of Utah, Salt Lake City, Utah; ‖‖Catholic University of Parana, Curitiba, Brazil; ***University of Pittsburgh, Pittsburgh, PA; †††MD Anderson Cancer Center, Houston, TX; ‡‡‡Northwestern University, Chicago, IL; and §§§Medtronic Spinal and Biologics, Memphis, TN.

Acknowledgment date: December 17, 2007. Revision date: July 7, 2008. Acceptance date: September 21, 2008.

The manuscript submitted does not contain information about medical device(s)/drug(s).

Funds were received in support of this work. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

Supported by the Spine Oncology Study Group and funded by an educational/research grant from Medtronic Spinal and Biologics.

The Spine Oncology Study Group acknowledges prior work on this topic by Bradford L. Currier, MD, Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN.

Address correspondence and reprint request to Charles G. Fisher, MD, Vancouver General Hospital, 2733 Heather Street, D6 Heather Pavilion, VGH, Vancouver, BC V5Z 3J5; E-mail: Charles.Fisher@vch.ca

Despite advances in the treatment of primary spinal tumors in recent years, there remains a lack of consensus with respect to the feasibility of oncologically appropriate surgical treatment and the selection of the optimal surgical approach. These tumors are relatively rare, comprising 11% of all primary musculoskeletal tumors and 4.2% of all spine tumors.[1,2] Of all primary spine tumors only 6% are malignant,[3] but it is the malignant tumors that present the greatest therapeutic challenges. The uniqueness of these tumors has lead to a lack of evidence-based standards and the potential immensity of these cases has resulted in varied surgical management based on individual surgeon's experience and preference. In addition, most surgical options carry significant morbidity and consume vast resources. In contrast, there is emerging evidence that incomplete or oncologically inappropriate resection increases local recurrence rate and decreases overall survival.[4–15]

Table 1. Modified Articulation of Enneking Stages With Surgical Margins (Boriani *et al*)

Enneking Stages	Margin for Control
1	No management unless for decompression or stabilization
2	Intralesional excision ± local adjuvants
3	Marginal *en bloc* excision
IA	Wide *en bloc* excision
IB	Wide *en bloc* excision
IIA	Wide *en bloc* excision + effective adjuvants
IIB	Wide *en bloc* excision + effective adjuvants
IIIA	Palliative
IIIB	Palliative

There are currently 2 staging systems used to classify spine tumors and to make better-informed treatment planning. The Enneking classification was introduced in the 1980s for the management of appendicular musculoskeletal tumors.[16,17] The tumor is staged by its biologic aggressiveness, anatomic extent, and presence of metastasis. The Enneking classification appears valid in predicting the prognosis and guides the choice of surgical margins in patients with primary tumors of extremities (Table 1); however, the adoption of this classification in the management of primary spine tumors entails complexities not encountered in the appendicular skeleton. It does not account for the existence of a continuous epidural compartment, the neurologic implication of sacrificing the spinal cord and roots, and the need for restoring spinal stability.

The Weinstein-Boriani-Biagini (WBB) classification[5,18] was devised to stage the spinal tumor while recognizing the unique anatomic complexity of the spine. It provides guidelines as to the feasibility and type of necessary surgical resection. The fundamental concept of this system is to ensure sparing of spinal cord without compromising the surgical tumor margins (Table 2).[18]

The application of the Enneking and WBB classification to primary spine tumors has been studied and appears safe, feasible and to improve disease control and survival in patients.[5,15] However, before these classifications can be definitively validated and their generalizability assessed, their reliability must be determined. A reliable classification means that there should be adequate agreement among the treating clinicians on staging the same patient (interobserver reliability) and the same staging results should be repeatable by the same clinician in different settings (intraobserver reliability). If the reliability and subsequent validity of one or both of these classifications is established, it will provide an evidence-based standardized approach to treat and study these uncommon, potentially lethal tumors. Therefore the purpose of the study is to evaluate the intraobserver and interobserver reliability of both the Enneking and WBB classifications for the management of primary tumors of the spine.

Table 2. Articulation of WBB Stages With Surgical Procedures

Radiating Zone	Procedure
4–8 or 5–9	Vertebrectomy (double approach)
2–5 or 7–11	Sagittal resection (double approach)
10–3	Posterior arch resection (posterior approach)

There are 3 major methods for performing *en bloc* resections: (1) vertebrectomy; (2) sagittal resection; and (3) posterior arch resection.

■ Materials and Methods

Case Selection and Evaluation

A representative sample of 15 cases of primary spinal tumors was selected from a prospectively collected, fully relational spine database at a tertiary care referral center. The cases reflected different anatomic levels, extent of involvement, and a variety of primary pathology and biologic behavior. More malignant then benign were selected because of the complex management issues. Chondrosarcoma, the most common primary malignant tumor (15% of all chondrosarcomas) was the most represented.[19] Preoperative computed tomography (CT) and magnetic resonance imaging (MRI) scans, demographic data (age and gender), clinical information, and pathologic data (tumor histopathology and distant metastasis) were included in 2 sets of handouts and compact discs (CDs) (Figure 1). The handouts also contained instructions and a case example that clarified the surgical terminology and the methodology of the Enneking and WBB classifications, as well as the answer sheets. The second set of handouts and CDs contained the same cases as in the first set, but in a different order.

These materials were sent to 18 experienced spine surgeons from the Spine Oncology Study Group, an international group of orthopedic and neurosurgical spine surgeons dedicated to the study and advancement of spine oncology. Using the clinical information and imaging studies, each observer was asked to stage the tumor cases using the 2 classifications by marking boxes on the answer sheet that best described the Enneking tumor pathology, Enneking tumor extent, Enneking metastatic status, Enneking stage, WBB sector, and WBB layers. Based on the staging, each observer was asked to formulate a treatment plan by marking boxes on the answer sheet that best described the Enneking surgical margins and WBB approaches to surgical resection (Figure 2).

Once completed, the results were sent to an independent central study coordinator. The identical evaluation was then conducted 4 weeks later with the cases in a different order to limit recall bias.

Staging

Although a detailed discussion of the staging is beyond the scope of this manuscript, a brief description is worthwhile. The Enneking classification is based on the interrelationship of the biologic grade (G), the local extent of the tumor (T), and the presence of metastases (M). The tumors are divided into 3 grades according to their biologic behavior, with G_0 denoting benign tumors, G_1 low-grade malignant tumors, and G_2 high-grade malignant tumors. Benign tumors include giant cell tumor, osteoid osteoma, osteoblastoma, osteochondroma, chondroma, and chondroblastoma. Malignant tumors include chondrosarcoma, Ewing sarcoma, and osteosarcoma. The local extent of the tumor varies from intracapsular (T_0), through intracompartmental (T_1), to extracompartmental (T_2). Metastases may be absent (M_0) or

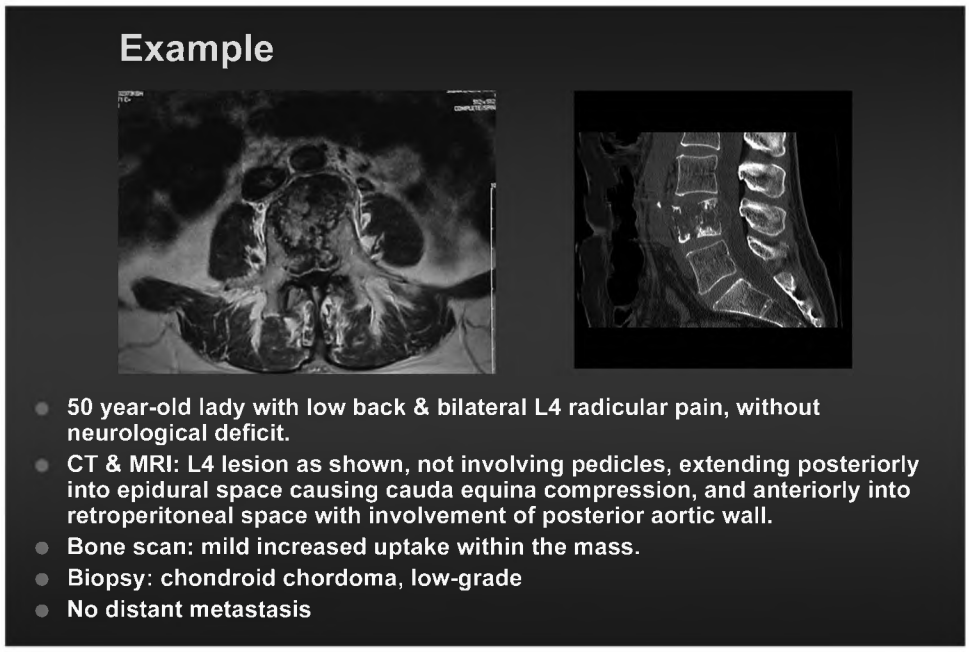

Figure 1. An example of cases compiled and sent to the raters for review.

present (M_1). These 3 factors combine to give the Enneking stages (Table 1).

Using the surgical principles dictated by the Enneking classification, the recommended surgical margins can be intralesional (with plane of dissections within the lesions), marginal (dissection within the reactive zone or pseudocapsule), wide (dissection beyond the reactive zone through the normal tissue), and radical (extracompartmental dissection) (Table 1). In spine surgery, radical margins are not feasible with the theoretical exception of a stage I/IIA tumor totally confined to the vertebra (no epidural disease) where a complete resection including the spinal cord is performed. This is attributed to the fact that the epidural space is 1 continuous compartment extending from the occiput to the sacrum.

The WBB staging system divides the spine in the axial plane into 12 equally radiating zones centered about the spinal canal and numbered from 1 to 12 in a clockwise fashion, with zone 1 and 12 located at the left and right side of spinous process, respectively. The tumor is further divided into 5 concentric layers centered about the dural sac and ranging from layers A (extraosseous soft tissues), B (intraosseous superficial), C (intraosseous deep), D (extraosseous extradural), to E (extraosseous intradural). Finally, the longitudinal extent of the tumor is recorded as segments of vertebrae involved. Based on the WBB stages, Boriani *et al* proposed indications for surgical procedures based on their experience with 29 patients (Table 2).[5]

Statistical Analysis

The interobserver and intraobserver reliability of the variables of Enneking grade, tumor and metastasis categories, Enneking stage, and Enneking-recommended surgical margin, as well as the WBB zones, layers, and recommended surgical procedure were assessed among surgeon raters *via* the calculation percent agreement/proportional overlap. To take into account the fact that observers will sometimes agree or disagree simply by chance, the κ statistic was applied to each variable (Table 3). The methods of assessment were determined by the type of variables analyzed.

For the mutually exclusive categorical variables of Enneking grade, tumor extent, metastasis, stage, and recommended surgical margin, the percent agreement among raters was calculated. Fleiss κ statistic was applied to all interobserver assessments except Enneking metastasis, which is a binary categorical variable.[20] Therefore, Cohen κ statistic was calculated for the assessment of the interobserver reliability of Enneking metastasis.[21] Intraobserver reliability for these mutually exclusive categorical variables was assessed *via* the calculation of the percent agreement and Cohen κ statistic. All assessments for mutually exclusive categorical variables were calculated using SPSS version15.0 (SPSS Inc., Chicago, IL).

WBB zones and layers are nonmutually exclusive categorical variables, therefore both the interobserver and intraobserver reliability of these data fields were assessed *via* the calculation of proportional overlap and Mezzich κ statistic, using Microsoft Excel 2003.[22,23]

■ Results

Interobserver Reliability of Enneking Classification

The interobserver reliability of Enneking staging system has a Fleiss κ statistics of 0.57 (moderate level of agreement) (Table 4). The analyses of Enneking subcategories revealed interobserver reliability of 0.82 and 0.22 (Fleiss κ statistics) for the fields of Enneking grade and Enneking tumor extent, respectively. The interobserver reliability of Enneking metastasis subcategory, which is a binary categorical variable, is calculated with Cohen κ statistic and yielded a κ coefficient of 0.00. The Fleiss κ coefficient of interobserver reliability of Enneking-recommended surgical margin is 0.47 (moderate level of agreement).

Interobserver Reliability of WBB Staging

Using Mezzich κ statistics, the κ coefficients of interobserver reliability for WBB zones and WBB layers are 0.31 (fair level of agreement) and 0.58 (moderate level of agreement), respectively (Table 5). The WBB-recommended

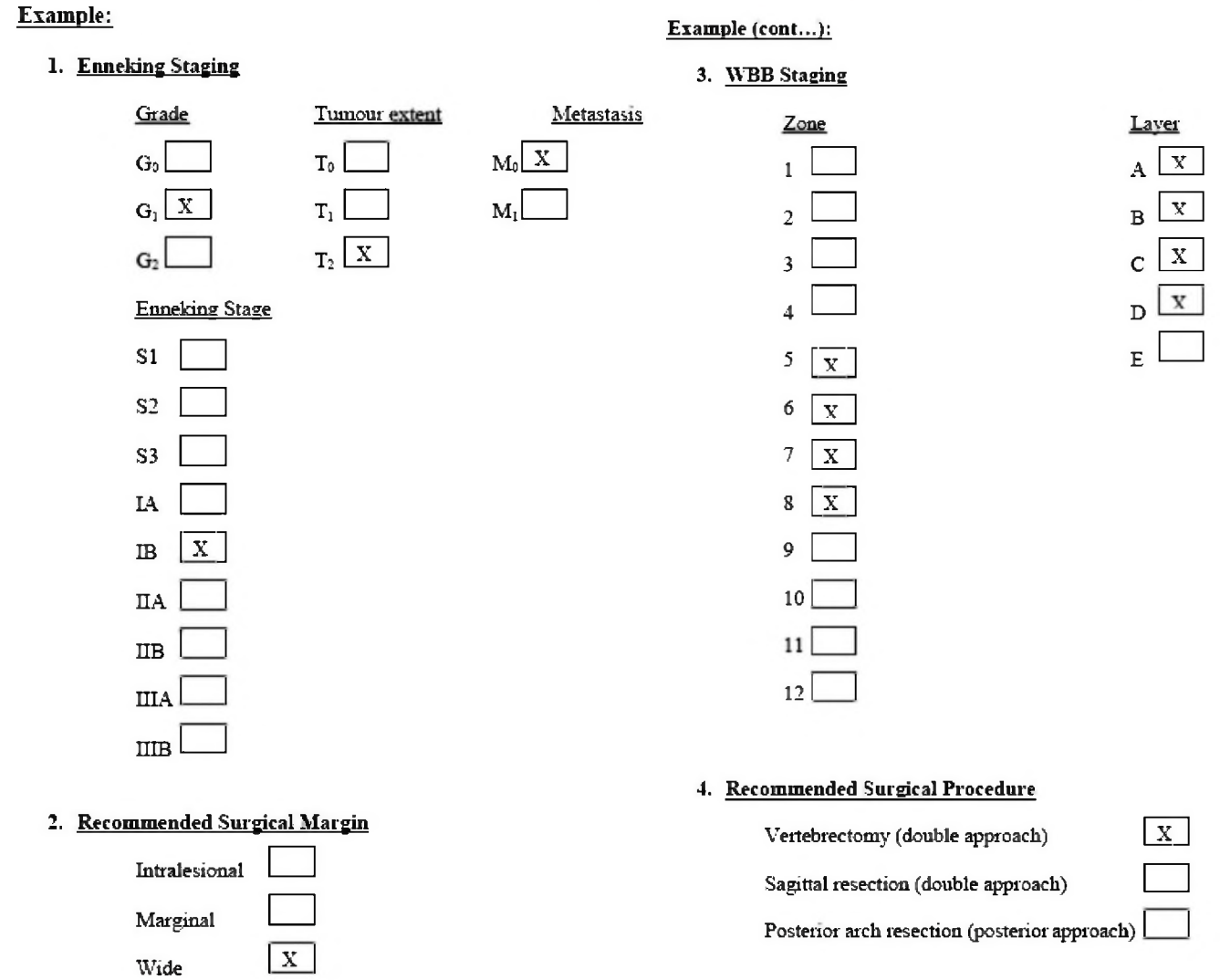

Figure 2. An example of handouts with categories of variables to be reviewed and scored by the individual raters.

surgical procedures have a Fleiss κ values of 0.54 (moderate level of agreement).

Intraobserver Reliability of Enneking Classification

The intraobserver reliability of Enneking staging system has a Fleiss κ statistics of 0.82 (near-perfect level of agreement) (Table 4). The analyses of Enneking subcategories revealed intraobserver reliability of 0.97 and 0.53 (Fleiss κ statistics) for the fields of Enneking grade and Enneking tumor extent, respectively. The intraobserver reliability of Enneking metastasis subcategory, which is a binary categorical variable, is calculated with Cohen κ statistic and yielded a κ coefficient of 0.47. The Fleiss κ coefficient of interobserver reliability of Enneking-recommended surgical margin is 0.67 (substantial level of agreement).

Table 3. Percentage of Agreement at a Variety of κ Statistics Levels (Landis and Koch)

κ Value	Levels of Agreement
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
>0.80	Near perfect

Intraobserver Reliability of WBB Staging

Using Mezzich κ statistics, the κ coefficients of intraobserver reliability for WBB zones and WBB layers are 0.63 and 0.79, respectively (both have substantial level of agreement) (Table 5). The WBB-recommended surgical procedures have a Fleiss κ values of 0.79 (substantial level of agreement).

Benign Versus Malignant Tumors

A comparison of the above parameters between benign and malignant tumors showed similar results with the exception of "tumor grade," which were slight (0.001) in benign and near-perfect in malignant (0.803) (Tables 6, 7). In "tumor extent," benign was fair (0.271) and malignant was slight (0.077). The percent agreement/proportional overlap for all Enneking and Weinstein-

Table 4. Results of Reliability Analysis on the Enneking Staging System

	Interobserver Reliability			Intraobserver Reliability		
	N	κ	Percent Agreement	N	κ	Percent Agreement
Enneking grade	314	0.82	89.8%	258	0.97	98.1%
Enneking tumor extent	304	0.22	81.8%	254	0.53	88.9%
Enneking stage	314	0.57	72.3%	269	0.82	88.5%
Enneking-recommended surgical margin	314	0.47	77.6%	269	0.67	86.2%

Boriani-Biagini measures, stratified by tumor cell type, was calculated and is presented in Table 7.

■ Discussion

To our knowledge this is the first study to examine, in a large, international cohort of neurosurgical and orthopedic spine surgeons, the interobserver and intraobserver reliability of the Enneking and WBB classifications for primary tumors of the spine.

Although the Enneking classification is adapted from appendicular musculoskeletal oncology, its principles of applying appropriate tumor margin resection appear valid from the perspective of tumor recurrence and patient survival.[16,17] The WBB classification was introduced for staging and guiding the surgical resection of primary spinal tumors.[5,18] It addresses the shortcomings of the former classification by accounting for the presence of the epidural compartment, the neural tissues and unique anatomy of the spine. Its safety, feasibility, and preliminary criterion validity was demonstrated by Boriani *et al*[5] and Fisher *et al*[15]; however, both classifications have not been subjected to rigorous assessment of their reliability. Before further validation, their reliability should be established by demonstrating adequate agreement and repeatability across examiners (interobserver reliability) and within patient populations (intraobserver reliability).

We calculated the percentage of agreement as a way of gauging the agreement among different raters and the agreement by the same raters in different timing, for most categories except the WBB zones and layers. In the latter 2 categories, each rating consists of a range of zones and layers, which may not be in complete agreement, but overlaps in the zones and layers occupied by the main bulk of the tumors. We are interested in the extent of overlap (proportional overlap) in the zoning and layering of the tumor, which is the basis for decision on the "WBB-recommended surgical approaches." To account for the agreement that would be expected purely by chance, the κ coefficient (κ) is calculated using the various described methodology according to the types of variables assessed, as discussed in the methods section.

The calculated magnitude of κ coefficients ranges from 0.00 to 1.00. It is not simple to assign a definite interpretation to κ coefficients, because it is dependent on the prevalence, the number of categories, possible weighting, and the presence of bias. Landis and Koch proposed the following as standards for interpretation of the strength of reliability with the κ coefficient: 0.00 to 0.20 = slight agreement, 0.21 to 0.40 = fair agreement, 0.41 to 0.60 = moderate agreement, 0.61 to 0.80 = substantial agreement, and >0.80 = near-perfect agreement.[24] Similar formulations and adaptations exist with slightly different descriptors.[25] The minimum acceptable value of κ coefficient depends on the clinical context and the choice of such benchmark is inevitably arbitrary. In agreement with most published medical journals, we consider a κ coefficient of less than 0.41 to be clinically "unacceptable" for our reliability study on Enneking and WBB staging systems.[26]

Enneking Staging

We demonstrated moderate interobserver reliability (κ = 0.57 and 0.47, respectively) and substantial to near-perfect (κ = 0.82 and 0.67, respectively) intraobserver reliability for the Enneking stage and Enneking-recommended surgical margin. The results suggested that although a treating surgeon is more likely to be consistent in assigning the same Enneking stage and planning for the same resection tumor margin, the agreement among different treating surgeons are only "moderate." "Moderate agreement" is theoretically considered clinically acceptable ($\kappa \geq 0.41$), and is comparable with accepted classification systems in spinal trauma sur-

Table 5. Results of Reliability Analysis on the WBB Staging System

	Inter-Rater Reliability			Intra-Rater Reliability		
	N	κ	Proportional Overlap	N	κ	Proportional Overlap
WBB zones	3130	0.31	0.48	269	0.63	0.77
WBB layers	3110	0.58	0.74	266	0.79	0.88
	N	κ	Percent Agreement	N	κ	Percent Agreement
WBB recommended surgical procedure	309	0.54	75.10%	263	0.79	89.00%

Table 6. Stratified WBB and Enneking Reliability Analyses

Statistic	Enneking Grade	Enneking Tumor Extent	Enneking Metastasis	Enneking Stage	Enneking Recommended Surgical Margin	WBB Zones	WBB Layers	WBB-Recommended Surgical Procedure
Orthopedic surgeons only								
No. cases	149	148	149	149	149	146	146	146
Fleiss κ	0.892	0.485		0.814	0.724	0.319	0.725	0.548
Percent agreement	93.80%	91.56%	100.00%	88.81%	89.04%	0.483	0.841	74.30%
Agreement	Almost perfect	Moderate		Almost perfect	Substantial	Fair	Substantial	Moderate
Neurosurgeons only								
No. cases	165	165	165	165	165	164	164	164
Fleiss κ	0.759	0.097		0.383	0.289	0.294	0.488	0.571
Percent agreement	86.31%	75.15%	97.58%	58.67%	68.61%	0.455	0.656	78.30%
Agreement	Substantial	Slight		Fair	Fair	Fair	Moderate	Moderate
Benign tumors only								
No. cases	63	63	63	63	63	62	62	62
Fleiss κ	0.001	0.271		0.267	0.237	0.330	0.555	0.707
Percent agreement	82.54%	67.30%	98.41%	60.63%	57.5%	0.496	0.714	80.81%
Agreement	Slight	Fair		Fair	Fair	Moderate	Moderate	Substantial
Malignant tumors only								
No. cases	251	250	251	251	251	247	247	247
Fleiss κ	0.803	0.077		0.522	0.292	0.310	0.587	0.424
Percent agreement	91.56%	85.56%	99.21%	75.13%	82.7%	0.473	0.740	73.70%
Agreement	Almost Perfect	Slight		Moderate	Fair	Moderate	Moderate	Moderate

gery such as such as the Arbeitsgemeinschaft für Osteosynthesefragen (AO), Denis and Thoracolumbar Injury Severity Score (TLISS) classification systems.[27–31] It is important to realize however, that moderate agreement equates to an extra 41% to 60% greater agreement over that expected by chance. Therefore, there exist potentially chances of interobserver disagreement inherent in the Enneking staging and Enneking-recommended surgical margin.

The potential interobserver disagreement of Enneking classification appears to be mainly related to the Enneking tumor extent subcategory that has only a fair interobserver reliability ($\kappa = 0.22$), and the κ coefficient for its interobserver reliability is further lowered by the low prevalence paradox of Enneking Metastasis subcategory. The Enneking tumor extent subcategory assesses the raters' opinion if a tumor is intracapsular, intracompartmental, extracapsular, or has extracompartmental extensions based on MRI and CT scanning. The exact tumor margin and its respect for or invasion of anatomic barrier may not be reliably determined on the somewhat limited "representative images" provided in this study and is more likely to improve with a full series of imaging sequences in all planes. The logistical and feasibility issues to provide full imaging series to participating surgeons in this study was overwhelming and not felt to be necessary, as representative images biased toward a lower κ. Furthermore neurosurgical spine surgeons were less familiar with the Enneking classification and this probably further compromised interobserver reliability (Table 6).

The subcategory of Enneking metastasis was calculated to have a κ coefficient of 0.00 for interobserver reliability, despite its high percent agreement of 98.7%. This paradox of "high agreement-low κ" occurs because of the homogeneity of the cases (all no metastases–M_0). It is suggested that for analysis with a high agreement but a low κ, more emphasis should be laid on the percent agreement.[32–36] Far beyond the statistical explanation is the practical reality that through appropriate systemic staging (bone scan, chest/abdominal CT, positron emission tomography scan *etc.*) the presence of metastases is a clear yes or no reported to the surgeon. The only vari-

Table 7. Percent Agreement by Tumor Type

Histology	No. Cases	Enneking Tumor Extent	Enneking Stage	Enneking Surgical Margin	WBB Zones (Proportion Overlap)	WBB Layers (Proportion Overlap)	WBB Recommended Surgical Procedure
Chondrosarcoma	6	93.7%	81.8%	86.8%	0.413	0.716	82.1%
Giant cell tumor	2	83.8%	69.8%	63.8%	0.621	0.791	80.3%
Osteosarcoma	2	78.1%	66.2%	82.4%	0.560	0.846	47.6%
Ewing sarcoma	1	100.0%	100.0%	81.3%	0.438	0.871	66.3%
Hemangioendothelioma	1	67.6%	57.6%	73.3%	0.830	0.750	30.5%
Leiomyosarcoma	1	100.0%	81.3%	100.0%	0.434	0.623	30.5%
Osteoblastoma	1	41.1%	49.0%	46.2%	0.373	0.644	49.5%
Osteoid osteoma	1	34.3%	42.4%	44.8%	0.247	0.560	81.3%

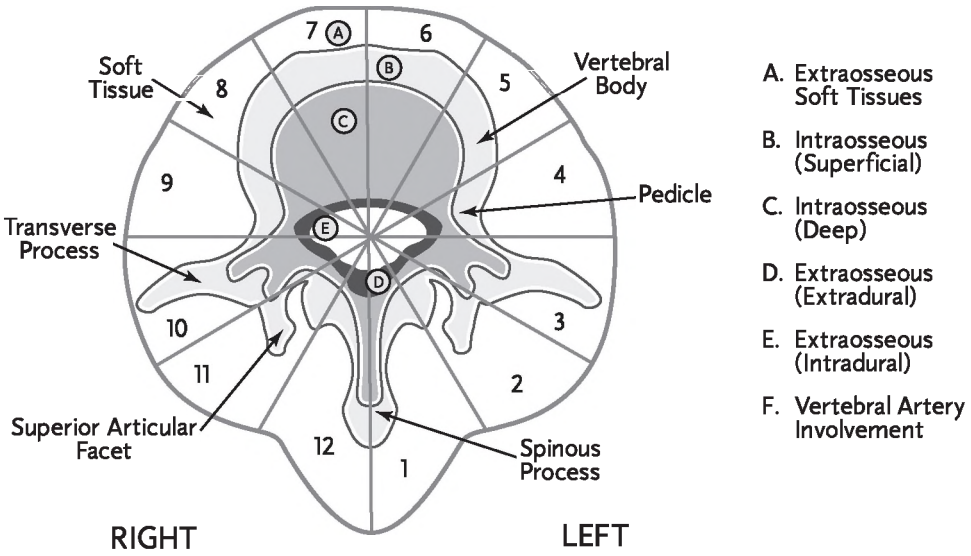

Figure 3. Modified WBB staging by consensus of the Spine Oncology Study Group. The diagram is now in the same orientation as in conventional CT scan and MRI images.

ability is related to the diagnostic tests used to determine the presence of metastases.

WBB Staging

Both categories of WBB zones and layers consist of nonmutually exclusive categorical variables. A tumor was defined as having a "range" of WBB zones and layers. The difference in a few zones or layers within the range of overlapping WBB stages assigned by different raters does not constitute absolute disagreement. Complete disagreement occurs only if there is no overlap between the "ranges" assigned. Of clinical importance is the quantifiable proportion of agreement, which is the proportional overlap of the rating. Cohen κ method, which is designed for mutually exclusive categorical variables by determining match/no match in all rating pairs, will result in a misleadingly lower κ coefficient. Instead, we reassessed WBB zones and layers categories using Mezzich method that calculates a proportional overlap among rating pairs, and showed the interobserver reliability to be fair (κ = 0.31) and moderate (κ = 0.58), respectively.[22,23] One potential contribution to the lower interobserver reliability is the diagrammatic zones of the WBB do not match in orientation to the conventional axial cuts of the CT and MRI. The intraobserver reliability was substantial (κ = 0.63 and 0.79, respectively).

In contrast, the κ coefficient for the WBB-recommended surgical procedures is determined by analyzing their percent agreement using Cohen κ method, which is designed for mutually exclusive categorical variables. It had moderate interobserver reliability (κ = 0.54) and substantial intraobserver reliability (κ = 0.79). Similar to the Enneking classification, the reliability of the WBB staging was mainly affected by the difficulty in the determination of the exact tumor margin and the invaded anatomic compartment from the "representative imaging studies." However, with the main bulk of tumor being discernible in its anatomic location, moderate agreement among surgeons in deciding the recommended surgical procedures is still possible.

Limitations

The lack of the ability to view the images in multiple slices in each plane may reduce the ability to mentally reconstruct a three-dimensional representation of the tumor, thus potentially diminishing the interobserver reliability in defining the Enneking tumor extent, and WBB zones and layers. To reduce observer bias and enhance generalizability, the raters were selected from experienced oncology spinal surgeons from different centers across North America and abroad. Although explanation and instructions were sent to the raters, it was assumed that they were experienced in and familiar with the 2 staging systems assessed and the principles of oncological spine surgery. The range and extent of their experience in these regards were not formally evaluated but when neurosurgical spine surgeons and orthopedic spine surgeons were analyzed separately the higher κs in the orthopedic group is probably due to a greater familiarity with the classification due to background and training in appendicular oncology.

The Spine Oncology Study Group proposes the following minor but key modification to the WBB system to make it more user-friendly and to enhance reliability: orientating the zones so they are consistent with conventional MRI and CT axial cuts (Figure 3).

This study shows that the intraobserver reliability for both Enneking and WBB classifications are substantial to near-perfect; however, the interobserver reliability was considered fair to moderate. Although primary spinal tumors are uncommon they represent an immense therapeutic challenge fraught with morbidity and mortality, and therefore demand a reliable, validated, evidence-based classification on which to base treatment and conduct future research. The reliability results from this study provide a sound foundation on which to prospectively assess reliability and the crite-

rion validity of these classifications so as to optimize patient management and clinical research from a true global perspective.

■ Key Points

- Reliability study of the Enneking and WBB staging systems for the evaluation and management of primary spinal tumors.
- Evaluation of interobserver and intraobserver reliability *via* case series review by an experienced, international group of spine surgeons.
- Moderate interobserver reliability and substantial and near-perfect intraobserver reliability for both the Enneking and WBB classification was noted in terms of staging and guidance for treatment.
- Further work is required to investigate the validity of the classifications.

References

1. Dahlin DC, Unni KK. *Bone Tumors: General Aspects and Data on 8,542 Cases*. 4th ed. Sprinfield, IL: Thomas, 1986.
2. Boriani S, Biagini R, De Iure F, et al. Primary bone tumors of the spine: a survey of the evaluation at treatment at the Instituto Ortopedico Rizzoli. *Orthopedics* 1995;18:993–1000.
3. Abdu WA, Provencher M. Primary bone and metastatic tumors of the cervical spine. *Magn Reson Imaging Clin North Am* 2000;4:299–320.
4. Bergh P, Gutenberg B, Meis-Kindblom LG. Prognostic factors and outcome of pelvic, sacral, and spinal chondrosarcomas: a center-based study of 69 cases. *Cancer* 2001;91:1201–12.
5. Boriani S, Biagini R, DeLure F. En bloc resections of bone tumors of the thoracolumbar spine. A preliminary report on 29 patients. *Spine* 1996;21: 1927–31.
6. Kaiser TE, Pritchard DJ, Unni KK. Clinicopathologic study of sacrococcygeal chordoma. *Cancer* 1984;53:2574–8.
7. Shives TC, Dahlin DC, Sim FH, et al. Osteosarcoma of the spine. *J Bone Joint Surg Am* 1986;68:660–8.
8. Shives TC, McLeod RA, Unni KK, et al. Chondrosarcoma of the spine. *J Bone Joint Surg Am* 1989;71:1158–65.
9. Spanier SS, Shuster JJ, Vander Greind RA. The effect of local extent of the tumor on prognosis in osteosarcoma. *J Bone Joint Surg Am* 1990;72:643–53.
10. Sundaresan N, DiGiacinto GV, Krol G, et al. Spondylectomy for malignant tumors of the spine. *J Clin Oncol* 1989;7:1485–91.
11. Sundaresan N, Rosen G, Huvos AG, et al. Combined treatment of osteosarcoma of the spine. *Neurosurgery* 1988;23:714–9.
12. Sundaresan N, Steinberger AA, Moore F, et al. Indications and results of combined anterior-posterior approaches for spine tumor surgery. *J Neurosurg* 1996;85:438–46.
13. Talac R, Yaszemski MJ, Currier BL, et al. Relationship between surgical margins and local recurrence in sarcomas of the spine. *Clin Orthop* 2002; 397:127–32.
14. Weinstein JN, McLain RF. Primary tumors of the spine. *Spine* 1987;12: 843–51.
15. Fisher CG, Keynan O, Boyd M, et al. The surgical management of primary tumors of the spine. Initial results of an ongoing prospective cohort study. *Spine* 2005;30:1899–908.
16. Enneking WF. A system of staging musculoskeletal neoplasms. *Clin Orthop* 1980;153:106–20.
17. Enneking WF, Spanier S, Goodman M. A system for the surgical staging of musculoskeletal sarcoma. *Clin Orthop* 1986;204:9–24.
18. Boriani S, Weinstein JN, Biagini R. Primary bone tumors of the spine. Terminology and surgical staging. *Spine* 1997;22:1036–44.
19. Campanacci M. *Bone and Soft Tissue Tumors*. New York, NY: Springer-Verlag Wien; 1990.
20. Fleiss JL. Measuring nominal scale agreement among many raters. *Psycho Bull* 1971;76:378–81.
21. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
22. Eccleston P, Werneke U, Armon K, et al. Accounting for overlap? An application of Mezzich's κ statistic to test interrater reliability of interview data on parental accident and emergency attendance. *J Advanced Nursing* 2001;33: 784–90.
23. Mezzich JE, Kraemer HC, Worthington DRL, et al. Assessment of agreement among several raters formulating multiple diagnoses. *J Psychiat Res* 1981; 16:29–39.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
25. Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res* 1998;7:301–17.
26. Sim J, Wright CC. The κ statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257–68.
27. Blauth M, Bastian L, Knop C, et al. Inter-observer reliability in the classification of thoraco-lumbar spinal injury. *Orthopade* 1999;28:662–81 [in German].
28. Denis F. The three-column spine and its significance in the classification of acute thoracolumbar spinal injuries. *Spine* 1983;8:817–31.
29. Patel AA, Vaccaro AR, Albert TJ, et al. The adoption of a new classification system. Time-dependent variation in interobserver reliability of the thoracolumbar injury severity score classification system. *Spine* 2007;32:E105–10.
30. Vaccaro AR, Zeiller SC, Hulbert RJ, et al. The thoracolumbar injury severity score: a proposed treatment algorithm. *J Spin Disord Tech* 2005;18:209–15.
31. Wood KB, Khanna G, Vaccaro AR, et al. Assessment of two thoracolumbar fracture classification systems as used by multiple surgeons. *J Bone Joint Surg* 2005;87-A:1423–9.
32. Assendelft WJJ, Bouter LM, Knipschild PG, et al. Reliability of lumbar spine radiograph reading by chiropractors. *Spine* 1997;22:1235–41.
33. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;43:551–8.
34. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–9.
35. Haas M. Statistical methodology for reliability studies. *J Manipulative Physiol Ther* 1991;14:119–32.
36. Viera AJ, Garrett JM. Understanding interobserver agreement: the κstatistic. *Fam Med* 2005;37:360–3.