Spinal Instability Neoplastic Score: An Analysis of Reliability and Validity From the Spine Oncology Study Group


ABSTRACT

Purpose
Standardized indications for treatment of tumor-related spinal instability are hampered by the lack of a valid and reliable classification system. The objective of this study was to determine the interobserver reliability, intraobserver reliability, and predictive validity of the Spinal Instability Neoplastic Score (SINS).

Methods
Clinical and radiographic data from 30 patients with spinal tumors were classified as stable, potentially unstable, and unstable by members of the Spine Oncology Study Group. The median category for each patient case (consensus opinion) was used as the gold standard for predictive validity testing. On two occasions at least 6 weeks apart, each rater also scored each patient using SINS. Each total score was converted into a three-category data field, with 0 to 6 as stable, 7 to 12 as potentially unstable, and 13 to 18 as unstable.

Results
The $\kappa$ statistics for interobserver reliability were 0.790, 0.841, 0.244, 0.456, 0.462, and 0.492 for the fields of location, pain, bone quality, alignment, vertebral body collapse, and posterolateral involvement, respectively. The $\kappa$ statistics for intraobserver reliability were 0.806, 0.859, 0.528, 0.614, 0.590, and 0.662 for the same respective fields. Intraclass correlation coefficients for inter- and intraobserver reliability of total SINS score were 0.846 (95% CI, 0.773 to 0.911) and 0.886 (95% CI, 0.868 to 0.902), respectively. The $\kappa$ statistic for predictive validity was 0.712 (95% CI, 0.676 to 0.766).

Conclusion
SINS demonstrated near-perfect inter- and intraobserver reliability in determining three clinically relevant categories of stability. The sensitivity and specificity of SINS for potentially unstable or unstable lesions were 95.7% and 79.5%, respectively.

J Clin Oncol 29:3072-3077. © 2011 by American Society of Clinical Oncology

INTRODUCTION

Spinal cord compression from epidural tumor is often discussed as an indication for operation. A prospective randomized trial has demonstrated superiority of surgery and radiation therapy compared with radiation alone in the treatment of high-grade spinal cord compression for solid tumors.1 Spinal instability is a separate indication for surgery2-7 or percutaneous cement augmentation,8,9 but it has not received the same degree of scrutiny in the literature as spinal cord compression. This paucity of data may reflect the controversy that exists regarding instability resulting from neoplastic destruction of spinal elements, as evidenced by a wide variety of criteria published in the literature2-8,16 and significant differences of opinion suggested by spine surgeons.2,12,13

The Spine Oncology Study Group (SOSG) defines spine instability as the “loss of spinal integrity as a result of a neoplastic process that is associated with movement-related pain, symptomatic or progressive deformity and/or neural compromise under physiological loads.”14 The development of a standard and valid classification with easily assigned radiographic and patient factors was championed to aid communication and appropriate referral between oncologists, radiologists, and spine surgeons and facilitate prompt, optimized treatment plans. Furthermore, a classification system could lead to a more consistent therapeutic approach among spine surgeons and aid in education and scientific study.
Spinal Instability Neoplastic Score

An evidence-based process using the best available literature and expert-opinion consensus was used to develop the Spine Instability Neoplastic Score (SINS; Table 1). In this classification system, tumor-related instability is assessed by adding together six individual component scores: spine location, pain, lesion bone quality, radiographic alignment, vertebral body collapse, and posterolateral involvement of the spinal elements. The minimum score is 0, and the maximum is 18. A score of 0 to 6 denotes stability, 7 to 12 denotes indeterminate (possibly impending) instability, and 13 to 18 denotes instability. A surgical consultation is recommended for patients with SINS scores greater than 7.14

Patient Case Selection and Evaluation

The SOSG is an international group of 30 spine oncology experts and thought leaders from North America, Europe, South America, and Asia who were asked to contribute patient case examples with imaging and clinical information for the purpose of testing SINS reliability and validity. A total of 50 de-identified patient cases were obtained. Patient cases that did not contain sufficient history or quality imaging were excluded. To obtain a SINS score, the history must include a description of pain, especially as it relates to patient movement. Imaging must include computed tomography (CT) scan or magnetic resonance imaging; however, if it is the latter, x-ray films (or preferably CT) are also required to determine bone lesion quality (ie, lytic, blastic, or mixed). In the case of multiple spinal lesions, contributors identified the specific lesion they intended for scoring. Thirty patient cases were chosen, with roughly equal representation of cervical, thoracic, and lumbar spinal levels as well as a broad range of neoplastic instability (Table 2). Patient cases were classified as stable, potentially unstable, or unstable on the basis of anonymous voting by panel of experts (consensus opinion). On the basis of preliminary analysis results and after further discussion among SOSG members, SINS was modified to improve reliability by simplifying the scoring method so that the minimum score in each category was 0. In addition, regions of the spine were defined more clearly: functional levels were occiput-C2, C7-T2, T11-L1, and L5-S1; mobile levels were C3-6 and L2-4; semi-rigid spine was T3-T10; and rigid spine was S2-S5. On the basis of data from the evidence-based reviews12,13 and expert consensus, SINS was also modified to include consideration of lesion bone quality and nonmechanical back pain.14

Statistical Analysis

Three statistical tests were used to assess inter- and intraobserver reliability. The intraclass correlation coefficient (ICC) was used to measure both inter- and intraobserver agreement for total SINS scores (two-way mixed effect model, in which people effects are random, and measures effects are fixed).17 For each of the six components of SINS (ie, location, pain, bone quality, radiographic alignment, vertebral body collapse, and posterolateral involvement), Fleiss’s κ for multiple raters was used to measure interobserver agreement, and Cohen’s κ was used to evaluate intraobserver agreement.18,19

RESULTS

Interobserver Reliability

The interobserver ICC reliability for total SINS score was 0.846 (95% CI, 0.773 to 0.911). The analysis of SINS components revealed

<table>
<thead>
<tr>
<th>Level</th>
<th>Stable</th>
<th>Potentially Unstable</th>
<th>Unstable</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cervical</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Thoracic</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Lumbar</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>30</td>
</tr>
</tbody>
</table>

NOTE. Final case series was selected to represent range of spinal levels and grades of stability. Stability was determined by anonymous voting by panel of experts (consensus opinion).

<table>
<thead>
<tr>
<th>Patient Cases</th>
<th>Stable</th>
<th>Potentially Unstable</th>
<th>Unstable</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cervical</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Thoracic</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Lumbar</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>30</td>
</tr>
</tbody>
</table>

NOTE. Final case series was selected to represent range of spinal levels and grades of stability. Stability was determined by anonymous voting by panel of experts (consensus opinion).
interobserver reliabilities (Fleiss $\kappa$) of 0.790, 0.841, 0.244, 0.456, 0.462, and 0.492 for the fields of location, pain, bone quality, alignment, vertebral body collapse, and posterolateral involvement, respectively (Table 4). The level of agreement was substantial for location, near perfect for pain, fair for bone quality, and moderate for the radiographic criteria of alignment, vertebral body collapse, and posterolateral involvement.

**Intraobserver Reliability**

The intraobserver ICC reliability for total SINS score was 0.886 (95% CI, 0.868 to 0.902). Analysis of SINS components revealed intraobserver reliabilities (Cohen’s $\kappa$) of 0.806, 0.859, 0.528, 0.614, 0.590, and 0.662 for the fields of location, pain, bone quality, alignment, vertebral body collapse, and posterolateral involvement, respectively (Table 4). The level of agreement was near perfect for location and pain, substantial for alignment and posterolateral involvement of the spine, and moderate for bone quality and vertebral body collapse.

**Validity**

The Cohen’s $\kappa$ statistic for agreement between collapsed SINS category and consensus score was 0.712 (95% CI, 0.676 to 0.766), which represents substantial agreement. Table 5 lists the scores determined by SINS raters, collapsed into the three categories (0 to 6, stable; 7 to 12, potentially unstable; and 13 to 18, unstable) and cross-tabulated with the a priori consensus opinion (gold standard). All unstable patient cases (by consensus opinion) obtained scores by SINS of 7 or greater, which suggests referral for assessment by a spine surgeon.

By analyzing results as negative for stable lesions (scores 0 to 6) and positive for potentially unstable or unstable lesions (scores $\geq 7$), SINS could be subjected to a binary classification test. Sensitivity was 563/588 = 95.7%, and specificity was 171/215 = 79.5%. The type II error (false-negative rate) was 4.3%, and the type I error (false-positive rate) was 20.5%.

**DISCUSSION**

A major goal in the treatment of metastatic disease is the preservation or restoration of spinal stability.1-16 Prediction of fracture risk in oncology is an important goal of classification systems. Mirels51 proposed a scoring system that is useful for predicting fractures in the peripheral skeleton. In the spine, criteria to define instability, which were originally developed for use in trauma, are not directly applicable in the setting of neoplasia.2 Before the development of SINS, there were no accepted evidence-based guidelines for the classification of spinal instability secondary to tumor.12-14 This significant void left oncologists, radiologists, and other health care providers who manage these patients without tools to assess this critical component of care. Patients who develop spinal instability have or are at high risk of having neurologic deficit, severe pain, and progressive deformity.1-11

SOSG envisions that SINS will allow clinicians from multiple disciplines who care for patients with spine tumors to make informed decisions about instability and when to refer patients for surgical assessment. SINS should also help spine surgeons make decisions regarding optimal management for specific patients who present with metastatic spine disease. In this initial analysis of reliability and predictive validity, we chose ourselves as the evaluators; however, future work is required to evaluate the performance of SINS among other groups of raters, particularly oncologists and radiologists.

It must be emphasized that the mechanical integrity of the spine is only one component of the evaluation process when considering surgery. If patients have neural deficit or if imaging reveals

### Table 3. Level of Agreement for $\kappa$ Statistic Levels

<table>
<thead>
<tr>
<th>$\kappa$ Value</th>
<th>Level of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00-0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>&gt; 0.80</td>
<td>Near perfect</td>
</tr>
</tbody>
</table>

**NOTE.** Data adapted.20

### Table 4. Reliability Analysis for Six Component Variables of SINS

<table>
<thead>
<tr>
<th>SINS Component</th>
<th>Location</th>
<th>Pain</th>
<th>Bone quality</th>
<th>Alignment</th>
<th>Vertebral body</th>
<th>Posterior involvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$</td>
<td>0.790</td>
<td>0.841</td>
<td>0.244</td>
<td>0.456</td>
<td>0.462</td>
<td>0.492</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.774-0.806</td>
<td>0.823-0.859</td>
<td>0.227-0.260</td>
<td>0.438-0.474</td>
<td>0.449-0.475</td>
<td>0.476-0.508</td>
</tr>
<tr>
<td>Level</td>
<td>Substantial</td>
<td>Near perfect</td>
<td>Fair</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Agreement %</td>
<td>86.3</td>
<td>91.0</td>
<td>68.1</td>
<td>71.6</td>
<td>59.7</td>
<td>66.5</td>
</tr>
</tbody>
</table>

**NOTE.** $P < .001$ for all. Abbreviation: SINS, Spinal Instability Neoplastic Score.
spinal cord compression (with or without neural deficit), surgery is often indicated regardless of SINS score.\textsuperscript{1} Other important considerations include tumor histology, response to nonoperative treatment, prognosis, patient medical fitness, and patient preference as part of informed choice.\textsuperscript{3,9} 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 and predictive validity of a classification system for tumor-related spinal instability.

We calculated percentage of agreement as a way of gauging agreement among different raters and agreement by the same raters in different timing. To account for the agreement that would be expected purely by chance, $\kappa$ coefficients and ICC\textsuperscript{18} were calculated according to types of variables assessed, as discussed in Methods.

It is not simple to assign a definite interpretation to $\kappa$ coefficients, because it is dependent on prevalence, number of categories, possible weighting, and presence of bias. Landis et al\textsuperscript{20} proposed standards for interpretation of the strength of reliability with the $\kappa$ coefficient (Table 3). Similar formulations and adaptations exist with slightly different descriptors.\textsuperscript{22}

The minimum acceptable value of $\kappa$ coefficient depends on clinical context and choice of such benchmark is inevitably arbitrary. Many medical journals consider a $\kappa$ coefficient of less than 0.41 to be clinically unacceptable for reliability.\textsuperscript{23} The inter- and intraobserver ICC reliability of final SINS categorization achieved near-perfect agreement. All subcategories had moderate to near-perfect agreement, with the exception of bone quality, which only had fair interobserver reliability and moderate intraobserver reliability. The decision to include this category was made based on strong biomechanical literature suggesting that the cross-sectional area of a defect combined with bone mineral density are excellent predictors of vertebral body failure and pathologic fracture risk.\textsuperscript{24-26} In those patients with metastatic lesions of the spine, low bone density is associated with greater fracture risk.\textsuperscript{27,28} Removing this component of SINS might improve reliability but would be detrimental to content validity. We may have been able to improve reliability by providing raters with multislice CT for every patient case.

The issue of validity is far more complex than that of reliability. Content and face validity of SINS were facilitated by integrating the best evidence provided by two systematic reviews with expert consensus from the SOSG.\textsuperscript{12,13} Quantitative scores were assigned based on the relative importance of particular factors gleaned from the literature and refined by expert consensus.\textsuperscript{14} One objective of the current study was an analysis of predictive validity. Obviously, the best method to determine the validity of a classification system is with prospective data; however, a preliminary assessment of the ability of SINS to predict instability was required before proceeding with a prospective protocol.

In medicine, a gold-standard test refers to a diagnostic test or benchmark regarded as definitive. The lack of a gold standard for measuring neoplastic spinal instability has hampered the development of a classification system.\textsuperscript{2} For this initial study, we considered two possible gold standards: contributor opinion and consensus opinion. Initially, we thought that the surgeon contributor of each patient case example would be the best to judge instability, because that person would presumably have knowledge of the outcome of the case. On the other hand, this would allow each contributor to impart significant personal bias. For example, if the outcome of a case was surgical stabilization, does that mean the spine was unstable? We believed using the median categorization of a panel of experts would bring us closer to a gold standard based on expert clinical opinion and the knowledge obtained by undertaking the creation of two systematic reviews.\textsuperscript{12,13} If consensus opinion and contributor opinion were equal gold standards, they should have a high level of agreement. In a previous abstract, we found only fair agreement between the three-tier category of stability (stable, potentially unstable, unstable) provided by the contributor and the consensus opinion (Cohen’s $\kappa$, 0.38).\textsuperscript{29}

SINS demonstrated high sensitivity (95.7%) for detecting lesions that were unstable or potentially unstable. This was based on combining the potentially unstable and unstable categories as disease positive in a binary classification test. All type II errors (4.3%) were potentially unstable patient cases categorized by SINS as stable. No unstable patient cases were incorrectly categorized as stable by SINS. Likewise, all type I error (20.5%) were stable but categorized by SINS as potentially unstable. In summary, for all raters, SINS did not classify any unstable patient cases as stable or vice versa. The positive predictive value of SINS cannot be calculated with the current data, because this value is directly proportional to prevalence of disease, which is not likely to be the case in a selected, retrospective case series.\textsuperscript{20}

Each patient case included a clinical history regarding pain, which is an important component of SINS. In the real world, history taking in this area may be more complex because of multiple sources of pain in patients with metastatic disease. We suspect this has artificially increased the interobserver reliability of pain categorization. Despite being provided the pattern of pain and location of the lesions, the panel still did not have perfect agreement on scores for pain and location, demonstrating some degree of imprecision with scoring and understanding of instructions.

Lack of the ability to view the images in multiple slices in each plane may have reduced the ability to determine the full extent of neoplastic involvement of the spine, thus potentially diminishing reliability of the imaging components of SINS. Bone quality in particular was difficult to score in some patient cases, because although magnetic resonance imaging was provided for every patient case, CT images were not always available. Although plain x-ray is helpful, the sensitivity of CT for assessing bony characteristics is greater.\textsuperscript{31}

Total SINS scores had near-perfect inter- and intraobserver reliability when collapsed into three clinically relevant assessments of tumor-related instability, which can be described as stability (scores 0 to 6), indeterminate (possibly impending) instability (7 to 12), and instability (13 to 18). Each component of SINS, with the exception of bone quality, demonstrated clinically acceptable reliability. The false-negative rate was low (4.3%), and all type II errors resulted from distinguishing stable and potentially unstable patient cases. No unstable patient cases were classified as stable by any SINS rater or vice versa.

Although SINS was developed in part to allow non–spine surgeons to make informed decisions about spinal stability, the raters in this study were a panel of experts. Further determination of SINS reliability is required among other groups of raters, particularly oncologists and radiologists. The prospective application of SINS in different patient populations is required to better determine validity and assess the application of this tool in therapeutic decision making.
with a “U” are those for which no compensation was received; those relationships marked with a “C” were compensated. For a detailed description of the disclosure categories, or for more information about ASCO’s conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Employment or Leadership Position: David W. Polly, Scoliosis Research Society (U) Consultant or Advisory Role: Christopher I. Shaffrey, Medtronic (C), Biomet (C), DePuy Spine (C); Sigurd H. Berven, Medtronic (C), DePuy Spine (C), Stryker (C), Biomet (C), Orthovita (C); David W. Polly, Medtronic (C); Shane Burch, Medtronic (C) Stock Ownership: Ziya L. Gokaslan, Spinle Kinetics, US Spine Honoraria: Christopher I. Shaffrey, Stryker; Sigurd H. Berven, Medtronic, Stryker, Biomet, Orthovita; James S. Harrop, Medtronic; Shane Burch, Medtronic. Research Funding: Christopher I. Shaffrey, National Institutes of Health, Department of Defense, AOSpine; Sigurd H. Berven, Medtronic, AOspine, Orthopaedic Research and Education Foundation; David W. Polly, Department of Defense, Scoliosis Research Society, Pediatric Orthopaedic Society of North America Expert Testimony: None Other Remuneration: Mark B. Dekutoski, Medtronic, Broadwater Associates, Synthes, DePuy Spine, Stryker


Manuscript writing: All authors

Final approval of manuscript: All authors

REFERENCES


AUTHOR CONTRIBUTIONS


Manuscript writing: All authors

Final approval of manuscript: All authors
Affiliations

Daryl R. Fourney and Evan M. Frangou, University of Saskatchewan, Royal University Hospital, Saskatoon, Saskatchewan; Christian P. DiPaola and Charles G. Fisher, University of British Columbia, Vancouver, British Columbia; Michael G. Fehlings, University of Toronto, Toronto, Ontario, Canada; Timothy C. Ryken, Iowa Spine and Brain Institute, Waterloo, IA; Christopher I. Shaffrey, University of Virginia Medical Center, Charlottesville, VA; Sigurd H. Berven, Shane Burch, and Dean Chou, University of California at San Francisco, San Francisco, CA; Mark H. Bilsky and Yoshiya Yamada, Memorial Sloan-Kettering Cancer Center; Narayan Sundaresan, Mount Sinai Hospital, New York, NY; James S. Harrop, Thomas Jefferson University, Philadelphia; Peter C. Gerszten, University of Pittsburgh Medical Center, Pittsburgh, PA; Stefano Boriani, Ospedale Maggiore, Ausl Bologna, Bologna; Roberto Biagini, Regina Elena Institute, Rome, Italy; Meic H. Schmidt, Huntsman Cancer Institute and Clinical Neurosciences Center, University of Utah, Salt Lake City, UT; David W. Polly, University of Minnesota, Minneapolis; Mark B. Dekutoski, Peter S. Rose, and Scott H. Okuno, Mayo Clinic, Rochester, MN; Aruna Ganju, Northwestern University, Chicago, IL; Ziya L. Gokaslan and Daniel M. Sciubba, Johns Hopkins University, Baltimore, MD; Michael W. Groff, Beth Israel Deaconess Medical Center; Norbert J. Liebsch, Massachusetts General Hospital, Boston, MA; Ehud Mendel, The Ohio State University, Columbus, OH; Shreyaskumar Patel and Laurence D. Rhines, The University of Texas MD Anderson Cancer Center, Houston, TX; Katsuro Tomita, Kanazawa University School of Medicine, Kanazawa, Japan; Peter P. Varga, National Center for Spinal Disorders, Budapest, Hungary; Luiz R. Vialle, Catholic University of Paraná, Curitiba, Paraná, Brazil; and Frank D. Vrionis, H. Lee Moffitt Cancer Center and Research Institute and University of South Florida College of Medicine, Tampa, FL.