Interobserver and intraobserver variation in the assessment of the healing of tibial fractures after intramedullary fixation


From St Michael’s Hospital, Sunnybrook Health Sciences Centre, and the University of Toronto, Toronto and McMaster University, Hamilton, Canada

The reliability of the radiological assessment of the healing of tibial fractures remains undetermined. We examined the inter- and intraobserver agreement of the healing of such fractures among four orthopaedic trauma surgeons who, on two separate occasions eight weeks apart, independently assessed the radiographs of 30 patients with fractures of the tibial shaft which had been treated by intramedullary fixation. The radiographs were selected from a database to represent fractures at various stages of healing. For each radiograph, the surgeon scored the degree of union, quantified the number of cortices bridged by callus or with a visible fracture line, described the extent and quality of the callus, and provided an overall rating of healing.

The interobserver chance-corrected agreement using a quadratically weighted kappa (k) statistic in which values of 0.61 to 0.80 represented substantial agreement were as follows: radiological union scale (k = 0.60); number of cortices bridged by callus (k = 0.75); number of cortices with a visible fracture line (k = 0.70); the extent of the callus (k = 0.57); and general impression of fracture healing (k = 0.67). The intraobserver agreement of the overall impression of healing (k = 0.89) and the number of cortices bridged by callus (k = 0.82) or with a visible fracture line (k = 0.83) was almost perfect.

There are no validated scales which allow surgeons to grade fracture healing radiologically. Among those examined, the number of cortices bridged by bone appears to be a reliable, and easily measured radiological variable to assess the healing of fractures after intramedullary fixation.

Received 17 May 2000; Accepted after revision 10 July 2001

In order to be consistently useful, the interpretation of the healing of fractures from radiographs must be reproducible. The assessment of the healing of tibial fractures relies on serial radiological examinations with clinical correlation. Sarmento et al specified that for successful union in the tibia the following criteria must be observed: 1) the ability of the patient to bear weight painlessly; 2) the absence of clinically detectable movement at the site of the fracture; and 3) visible bridging callus across the fracture on plain radiography. Techniques other than plain radiography, such as radionuclide imaging, CT, ultrasound, and resonant frequency analysis show potential for quantifying healing in the experimental setting, but are costly and cumbersome in their clinical application.

Despite the large number of studies of tibial fractures which have used healing as an outcome measure, there is disagreement on the precise definition of radiological union. Often, clinical investigations list non-specific criteria, such as the formation of callus or the absence of a fracture gap, for healing on follow-up radiographs. Several authors have defined radiological union as the presence of bridging callus on at least two views. Elaborate methods, such as that of Hammer, Hammerby and Lindholm, grade the healing of the fractures on the presence of callus, the bridging of callus, and the presence or absence of lucent fracture lines. Tower et al devised a similar scoring system, but also included periosteal callus. Similarly, Lane and Sandhu assigned scores based on the formation of bone, fracture lines, and remodelling. Although each method gives a quantitative assessment, none has been evaluated for intra- and interobserver reliability.

Intramedullary fixation can make the decisions as to
when tibial fractures are healed even more difficult. Clinical examination becomes less useful since stiffness, as assessed by movement at the site of the fracture, is often not detectable in the presence of an intramedullary nail which can obscure bony detail on the radiograph and interfere with the assessment of callus and the fracture line. The physician usually has only four cortices (without the intervening diaphyseal bone) on two orthogonal radiographs to evaluate the progression of healing of the fracture.

To our knowledge, there are no published studies which examine agreement in the radiological assessment of the healing of tibial fractures. We have therefore studied inter- and intraobserver agreement among surgeons in their radiological assessment of the healing of a tibial fracture after intramedullary nailing.

Patients and Methods

Eligible patients included those who had: 1) diaphyseal fractures of the tibia; 2) treatment with statically locked intramedullary nails; and 3) a complete set of anteroposterior (AP) and lateral radiographs six months after operation. We excluded patients with multiple fractures of the tibia. We chose randomly from our institution’s trauma database 30 sets of AP and lateral radiographs of fractures of the tibial shaft treated by intramedullary fixation. Our hypothesis was that radiographs from a variety of patients with closed and open fractures at six months would represent a distribution in which 20% of the fractures had healed (no visible fracture lines), 20% had not healed (obvious fracture lines), and 60% were in an indeterminate stage. After selection we confirmed that the radiographs represented fractures at various stages of healing.

Three orthopaedic trauma surgeons of similar experience and an orthopaedic trauma fellow independently viewed the radiographs on two separate occasions at a mean of eight weeks apart. They were not involved in the selection of the radiographs. Each surgeon was blinded to the others’ interpretations of the films. The coded radiographs included no patient identifiers. One of us (DBW or MB) was present with each surgeon at the initial and eight-week assessment to ensure the standardisation of the study protocol such as the randomised assessment of radiographs and the blinding of the assessors to the patients’ history and the age of the fracture.

Each surgeon marked each of the 30 sets of radiographs using a scoring sheet which included the criteria proposed by Hammer et al.\textsuperscript{15} (Table I). The observer assigned the film a grade from 1 (corresponding to a healed fracture) to 5 (corresponding to a fracture which is not healed) based on the amount and quality of callus as well as the presence or absence of a fracture line. Each surgeon then documented his overall rating of fracture healing (healed, probably healed, indeterminate, probably not healed, not healed), the extent and quality of callus (trace, moderate, massive and disorganised, massive and organised, homogeneous with remodelling), the number of cortices (out of four) with bridging callus, and the number of cortices with a visible fracture line.

The initial ratings of the four surgeons provided the basis for estimates of interobserver reliability. A second rating carried out at a mean of eight weeks (5 to 10) after the first provided the basis for estimates of intraobserver reliability. Since patients will be well served if different surgeons assessing the same patient arrive at similar conclusions about fracture healing, we focused on interobserver reliability as our primary assessment but we also examined intraobserver variability.

We chose chance-corrected agreement using a weighted kappa (κ) value with quadratic weights to assess reliability. The value ranged from +1, with perfect agreement, to -1, which corresponds to absolute disagreement. A value of 0 represents an agreement no better than that which occurs by chance alone. Since multiple observers were involved in this study, the κ value was calculated according to the method described by Fleiss.\textsuperscript{17} Interpretation of the values was carried out according to the guidelines proposed by Landis and Koch\textsuperscript{18} which suggest that values of 0 to 0.2 represent slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, and 0.61 to 0.80 substantial agreement. A value above 0.80 is considered to be almost perfect agreement. We calculated 95% confidence intervals (CI) for each κ value. For the correlation analysis

<table>
<thead>
<tr>
<th>Grade</th>
<th>Radiological assessment</th>
<th>Fracture line</th>
<th>Stage of union</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Homogeneous bone structure</td>
<td>Obliterated</td>
<td>Achieved</td>
</tr>
<tr>
<td>2</td>
<td>Massive. Bone trabeculae crossing fracture line</td>
<td>Barely discernible</td>
<td>Achieved</td>
</tr>
<tr>
<td>3</td>
<td>Apparent. Bridging of fracture line</td>
<td>Discernible</td>
<td>Uncertain</td>
</tr>
<tr>
<td>4</td>
<td>Trace. No bridging of fracture line</td>
<td>Distinct</td>
<td>Not achieved</td>
</tr>
<tr>
<td>5</td>
<td>No callus formation</td>
<td>Distinct</td>
<td>Not achieved</td>
</tr>
</tbody>
</table>

\textsuperscript{15}D. B. WHELAN, M. BHANDARI, M. D. MCKEE, G. H. GUYATT, H. J. KREDER, D. STEPHEN, E. H. SCHEMITSCH

THE JOURNAL OF BONE AND JOINT SURGERY
Results

According to the criteria of Hammer et al.\textsuperscript{15} the surgeons achieved moderate overall agreement ($\kappa = 0.6$, 95% CI 0.52 to 0.68). Those grades in the scale which corresponded to relatively indeterminate states of healing, agreed less well ($\kappa = 0.05$ to 0.31). The surgeons’ general assessment yielded similar reliability ($\kappa = 0.65$, 95% CI 0.59 to 0.75).

Agreement between observers was highest for the determination of the number of cortices bridged by callus ($\kappa = 0.75$, 95% CI 0.61 to 0.89) and for the number of cortices showing a fracture line ($\kappa = 0.70$, 95% CI 0.56 to 0.84). Moderate agreement was seen in quantifying the extent and quality of fracture callus ($\kappa = 0.57$, 95% CI 0.43 to 0.71).

Point estimates for intraobserver agreement $\kappa$ values for the Hammer union scale, the overall impression of healing, and for the number of cortices bridged by callus were 0.76, 0.89 and 0.83, respectively (Table II).

The surgeons’ general assessment of fracture healing was highly correlated ($r = 0.95$, $p < 0.01$) with the responses of the radiological criteria of Hammer et al.\textsuperscript{15}

Discussion

The difficulty in evaluating the healing of tibial fractures from plain radiographs has been discussed.\textsuperscript{18,19} The definition of radiological union is inconsistently reported and the degree of sophistication by which it is determined is also highly variable. Some investigations define radiological union based on a single parameter such as the presence of bridging callus on at least two views.\textsuperscript{7,13,14} Several variables are consistently noted when considering radiological union, including the amount and quality of callus, the presence of bridging callus and of a fracture line, and the trabeculae which cross it.

By using the most conservative estimates of agreement (i.e., the lower boundary of the 95% CI) we found that four orthopaedic trauma surgeons were moderately consistent in assessing the extent of fracture healing irrespective of whether they used the criteria of Hammer et al\textsuperscript{15} (lower boundary $\kappa = 0.52$) or their general impression of fracture healing (lower boundary $\kappa = 0.59$). Substantial agreement occurred when surgeons recorded the number of cortices bridged by callus (lower boundary $\kappa = 0.61$). Although numerous investigations have considered bridging callus on two views as a definition of union, the number of cortices has not previously been used. The number of cortices bridged (from 0 to 4) lends itself well to the clinical setting of intramedullary fixation since the presence of the nail allows only the cortical detail to be assessed accurately. In studies which were designed to correlate the radiological appearance of fractures with the qualitative strength, Panjabi et al\textsuperscript{12,20} suggested that the best single predictor of strength is cortical continuity. Our findings of the high reliability of the assessment of the number of cortices bridged by callus supports its potential use in assessing fracture healing in the presence of an intramedullary nail.

Our finding that the surgeons’ general impression was as reliable as the radiological criteria of Hammer et al\textsuperscript{15} also raises questions about the need for any rating system beyond an overall assessment. The surgeons determined the Hammer criteria and the rating of cortices at the same time as they recorded their general impression. The overall rating may well have been influenced by the surgeons’ prior assessment of fracture healing using the Hammer scale; had they made this in isolation, agreement may have been much lower. Moreover, the significant correlation ($r = 0.95$) between the Hammer criteria and the overall assessment of fracture healing suggests that those factors presented in the Hammer criteria are similar to those used by surgeons in their assessment of healing. It also lends further support to the possibility that the responses on the Hammer scale influenced the assessment of healing.

Ultimately, any formal rating should not only provide additional information beyond an overall assessment but also, as Burstein\textsuperscript{21} has suggested, it should have a direct relationship to the outcome. Much work has been done in recent years on interobserver reliability and the intra-observer reproducibility of previously accepted classification systems for fractures.\textsuperscript{22-25} The rationale for this scrutiny is to ensure that the information which these classification systems provide is reproducible. Ultimately, such a classification can be used in guiding treatment or indicating a prognosis with relatively predictable results. The same principle of ensuring reproducibility and reliability can be applied to other aspects of the management of fractures, especially the evaluation of healing.

While our study involved trauma surgeons in a level-1 trauma centre, we believe the findings may be useful for non-trauma surgeons in a more general setting, as most fractures do not present to trauma centres and the most reliable indicator of union, the number of cortices bridged by callus, is also the most objective measure.

We recommend AP and lateral radiography of the tibia at each follow-up visit, with measurement of the number of cortices (0 to 4) bridged by callus, as the most reliable

| Table II. Overall intraobserver weighted $\kappa$ values for assessment of the healing of tibial fracture |
|---------------------------------------------------|-------------------|-----------------|
| Assessment                                        | Overall weighted $\kappa$ | 95% CI          |
| Hammer scale                                      | 0.76              | 0.60 to 0.92    |
| General impression                                | 0.89              | 0.79 to 0.99    |
| Cortices bridged by callus                        | 0.82              | 0.78 to 0.86    |
| Cortices with a fracture line                     | 0.83              | 0.77 to 0.89    |

VOL. 84-B, No. 1, JANUARY 2002
method for assessing the progression of the healing of fractures after intramedullary nailing.

This study was funded by a grant from the Father Sean O’Sullivan Research Centre, Hamilton, Ontario, Canada.

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

References


