Interobserver Reliability in Imaging-Based Fracture Union Assessment—Two Systematic Reviews

Paul W. L. Ten Berg, MD, PhD,* Rik B. J. Kraan, MD,† Sjoerd Jens, MD, PhD,† and Mario Maas, MD, PhD†

Objectives: (A) To investigate the specialty of observers involved in imaging-based assessment of bone fracture union in recent orthopaedic trials and (B) to provide a general overview of observer differences (in terms of interobserver reliability) in radiologic fracture union assessment that have been reported between surgeons and radiologists.

Data Sources: Two separate systematic reviews (A, B) of English-, German-, and French-language articles in MEDLINE and Embase databases using the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines were done, with the following time frames: (A) January 2016–August 2017 and (B) through November 2017.

Study Selection: (A) Clinical trials of surgical fracture treatment evaluating radiologic (non) union. (B) Interobserver studies reporting kappa-values or intraclass correlation coefficients as reliability coefficient for radiologic fracture union assessment. Inclusion criteria for both reviews were fractures of the appendicular skeleton and the use of radiographs or computed tomography.

Data Extraction: Data were independently retrieved by 2 reviewers.

Data Synthesis: Descriptive statistics and percentages were reported.

Results: (A) Forty-eight trials were included, whereof 33 (68%) did not report the observer’s specialty. Six trials (13%) reported surgeon observers only, and 6 (13%) reported radiologist observers only. The median number of observers is 1 (interquartile range, 1–2). (B) Thirty-one interobserver studies were included, whereof 11 (35%) included at least 1 surgeon and 1 radiologist. Interobserver reliability varied considerably across the various fracture types studied and outcome scale used and was often unsatisfactory (kappa or intraclass correlation coefficients of <0.7).

Conclusions: In most trials providing observer’s characteristics, radiologic fracture union was either rated by 1 surgeon or 1 radiologist. As interobserver reliability can be unsatisfactory, we recommend surgeons and radiologists to further intensify collaboration and trials to include at least 2 observers and associated reliability statistics.

Key Words: fracture, union, nonunion, reliability, interobserver, radiograph, CT, imaging, review

(J Orthop Trauma 2020;34:e31–e38)

INTRODUCTION

In orthopaedics, a precise assessment of bone fracture union is important for both clinical and research purposes.1 In clinical practice, discrepancies between interpretations of fracture healing may influence treatment plan and prognosis, and it may also confuse patients who are able to look up their own records. In clinical research of fracture treatment, discrepancies between interpretations limit quality and generalizability of the study.

Imaging-based assessment is a critical part in the overall assessment of fracture union and is a frequently used end point in clinical research.2 The ability to discriminate between radiologic united and ununited fractures (ie, diagnostic accuracy3) is influenced by several factors including fracture characteristics, such as fracture location, and imaging characteristics, such as imaging modality. Common methods to assess bony union include the use of radiographs or computed tomography (CT).4

Due to its subjective nature, radiologic assessment of fracture union is observer dependent. Common observers are radiologists and surgeons—2 distinct specialists who have had different training programs5 and use a wide variation of fracture union criteria.2 This subjectiveness results in certain measurement errors between different observers whom rate the same image, affecting diagnostic interobserver reliability.6 Reliability coefficients [eg, kappa and intraclass correlation coefficient (ICC)] provide information about the ability of scores to discriminate between patients despite such measurement errors.7,8 Knowledge about observer details and associated interobserver reliability is therefore required to critically appraise clinical studies with radiologic fracture union as end point. The Consolidated Standards of Reporting Trials guidelines recommend nonpharmacological studies to specify who assessed outcomes and the number of assessors.9
The aim of this study was to investigate the reporting of observer characteristics in recent orthopaedic literature of clinical trials on surgical fracture treatment, with special focus on observer specialty. Our second aim was to provide a general overview of observer differences (in terms of interobserver reliability) in radiologic fracture union assessment that have been reported between surgeons and radiologists. To this end, we conducted 2 separate systematic reviews (A and B) focusing on fractures of the appendicular skeleton.

METHODS

We conducted 2 separate systematic reviews of (A) orthopaedic trials on surgical fracture treatment and (B) interobserver studies on radiologic fracture union assessment. The search, data extraction, and quality assessment were performed by 2 independent reviewers (P.T.B., R.K.), according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines. Disagreements were resolved by consensus-based discussion or a third reviewer (S.J.).

Selection Criteria

Regarding review for (A), inclusion criteria were (1) randomized clinical studies [randomized controlled trials (RCTs)] on bone fracture treatment to promote fracture healing, with surgical fracture fixation in at least one treatment group, and (2) use of radiologic outcome including postoperative fracture (non) union. We excluded (1) nonrandomized studies, (2) diagnostic studies, (3) studies with focus on wound care, infection, pain control, thromboprophylaxis, and peri- or postoperative management, and (4) follow-up studies based on previously published trials.

Regarding review for (B), inclusion criteria were (1) interobserver studies on radiologic fracture healing assessment, (2) inclusion of at least 2 observers, and (3) reporting on ICC or kappa statistics. For diagnostic reliability of studies including multiple observers, it is recommended to analyze continuous data with ICC and categorical data with the kappa statistic. We excluded interobserver studies focusing on fracture classifications.

Inclusion criteria of both reviews were (1) clinical studies, (2) fractures of appendicular skeleton, (3) use of standard radiographs or computed tomography to assess (non) union, (4) full-length articles, and (5) English-, German-, and French-language articles.

Search and Selection

We performed a systematic search using Medline (via PubMed) and Embase (via Ovid) databases in August 2017. An appendix shows the search strategies (see Appendix, Supplemental Digital Content 1, http://links.lww.com/JOT/A820).

Regarding review of (A), to sensitively search for RCTs, we used validated search terms as recommended in the Cochrane Handbook. To scope the entire orthopaedic literature, we searched in 72 orthopaedic journals as listed by the Web of Science 2016. Additionally, we included 8 large general medicine journals, including Annals of Internal Medicine, British Medical Journal (BMJ), Journal of the American Medical Association (JAMA), Lancet, The New England Journal of Medicine, Archives of Internal Medicine, Canadian Medical Association Journal, and PLoS One. We restricted our search to recent trials published between January 2016 and August 2017.

Regarding review of (B), to our knowledge, no sensitive search for interobserver studies has been validated yet; alternatively, we established a search strategy including search terms used in previous systematic reviews of reliability studies. The search was not restricted to journal nor publication year. This latter search was updated in November 2017. After the search, references were uploaded in Endnote X7 (Thomas Reuter, London, United Kingdom), and duplicates were removed. Fig. 1 shows the flowcharts of selection process for both reviews.

Data Extraction

Regarding both reviews, the following study characteristics were independently retrieved by the 2 reviewers: (1) fracture location, (2) imaging modality, (3) number of observers and specialty, (4) radiologic (non) union criteria and outcome scale, (5) applied reliability statistics, and (6) reliability outcome for each individual radiologic (non) union parameter. For review of (A), we described sample size based on the number of treated patients; for review of (B), sample size was based on the number of evaluated images/ series rated separately.

Quality Assessment

Regarding review of (A), our objective was to investigate a specific quality parameter focusing on outcome assessment and potential observer bias in orthopaedic trials. Therefore, no further quality assessment was performed.

Regarding review of (B), to assess study quality, we used the Quality Appraisal of Reliability Studies checklist. This checklist includes 11 items that appraise methodological domains of reliability studies: (1) representativeness of subjects, (2) representativeness of observers, (3) blinding, (4) order of examination, (5) time interval between repeated measurements, (6) test application and interpretation, and (7) the statistical analysis of the data (see also Appendix, Supplemental Digital Content 1, http://links.lww.com/JOT/A820). Note: evaluation of interobserver reliability does not require repeated measurements, contrarily to intraobserver reliability.

Analysis

Data were collected in Microsoft Excel 2013 (Microsoft, Redmond, WA). Descriptive statistics and percentages were reported. Kappa and ICC values were similarly interpreted. Values above an arbitrary threshold of 0.7 were considered satisfactory, whereas values below 0.7 were considered unsatisfactory—a general rule of thumb often applied in clinical research. A meta-analysis was not performed due to heterogeneity of fracture locations and applied outcome scales.
RESULTS

Review (A): Orthopedic Trials on Surgical Fracture Treatment

Our first search resulted in a total of 48 RCTs of surgical fracture treatment having radiologic fracture union as outcome (Table 1). Hereof, 43 trials (90%) had a surgery versus surgery design, and 5 trials (10%) had surgery versus conservative. Most trials treated hip fractures including intertrochanteric and femoral neck fractures (N = 13; 27%). The median sample size was 74 patients [interquartile range (IQR): 49–112 patients; range, 8–1108 patients].

The majority of the trials (N = 29; 60%) did not provide any information about the number or the specialty of the observers assessing union. In total, 9 trials (19%) included radiologists as observer. Of the trials providing specific numbers of included observers, the median number was 1 (IQR: 1–2; range, 1–4). No trials reported information about measurement errors of fracture union ratings between the observers.

Eighteen trials (38%) did not report a radiologic definition of fracture union. The most frequently reported definition of union was bridging of the fracture at 3 cortices (Table 2). The most frequently reported definition of non-union was the absence of union at 6 months (7 trials), followed by absence of union at 9 months (4 trials), and at 4 months (1 trial). All but 2 trials (96%) reported final union/non-union outcome dichotomously. Nineteen trials (40%) also reported time to union.

Review (B): Interobserver Studies on Radiologic Fracture Union Assessment

Our second search resulted in a total of 31 interobserver studies on radiologic fracture union assessment (Table 1). Most studies included series of tibial fractures (N = 9; 31%) and used only radiographs to assess union (N = 26; 84%). The median sample size was 40 evaluated images/series (IQR, 32–82; range, 16–250). The median number of included observers was 4 (IQR, 3–6; range, 2–69). Most studies (N = 13 (42%)) included only surgeons. Eleven studies (35%) included at least 1 surgeon and 1 radiologist. Characteristics and outcomes of these latter 11 studies are reported in Table 4.

Across the 31 interobserver studies, a total of 73 fracture union outcome measures were evaluated. The most frequently reported outcome measure was overall fracture union (Table 3). The reliability of the 73 fracture union outcome measures were evaluated using ICC or kappa coefficients, and it ranged from 0.05 to 0.92. Forty-two of the 66 reported outcome measures of union resulted in unsatisfactory reliability (kappa or ICC, < 0.7). Table 4 provides an overview of the observer differences reported between at least 1 surgeon and 1 radiologist and shows unsatisfactory reliability (ICC or kappa, < 0.7) in the majority of the outcome
The Consolidated Standards of Reporting Trials guide- 
lines recommend nonpharmacological studies to specify who 
assessed outcomes and the number of assessors. Based on 
our first review, 68% of 48 orthopaedic trials on surgical 
fracture treatment failed to report details of the outcome ob- 
servers including their specialty. In the remaining trials, re- 
ported observers were surgeons only (13%), radiologists only 
(13%), and surgeons and radiologists (6%). In general, only 1 
outcome observer was involved per trial. Based on our second 
review, interobserver reliability varied considerably across 
the various fracture types studied and outcome scale used, 
and it was often unsatisfactory (kappa or ICC, < 0.7).

A systematic review including 123 clinical studies on 
treatment of long-bone fractures from 1996 to 2006 showed 
that 74% of the studies failed to report details of outcome 
observers. In the remaining studies, the reported observers 
were surgeons (19%), radiologists (2%), and surgeons and 
radiologists (4%). Combined with our results, radiologists 
seem to be increasingly involved in orthopaedic research, 
although the large number of studies failing to report observer 
details remains fairly the same after more than 10 years. 
Another systematic review from 2008 including 92 clinical 
studies showed a wide variation in definitions of radiologic 
fracture union. This is comparable with our findings and sug-
gests a lack of consensus among specialists. In addition, the 
latter review found only 2 studies reporting on measurement 
errors between observers, and it therefore recommended future 
trials to include reliability statistics. Despite 
this recommendation, in the current review, we found no trials 
reporting on kappa or ICC coefficients. Thus, after 10 years, 
providing reliability statistics is still not common practice in 
orthopaedic trials on fracture healing treatment. This may 
limit deriving conclusions from the observed findings, 
especially for time to union, which requires reliable measurements 
to distinguish between established and unestablished union 
during the continuum of fracture healing. Regarding the def-
inition of nonunion, based on our review, most studies used 
a 6- or 9-month cutoff criterion. A survey study among 577 
orthopaedic surgeons found a larger variation ranging from 2 
to 12 months of time interval after injury before diagnosing 
nonunion. Perhaps, most importantly, our results demonstrate 

**DISCUSSION**

The Consolidated Standards of Reporting Trials guide-
lines recommend nonpharmacological studies to specify who 
assessed outcomes and the number of assessors. Based on 

**TABLE 2. Results of 48 Orthopaedic Trials (Review A): 
Reported Definitions of Radiologic Fracture Union**

<table>
<thead>
<tr>
<th>Union Definition</th>
<th>RCTs, N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>18 (38)</td>
</tr>
<tr>
<td>Bridging of fracture at 3 cortices</td>
<td>12 (25)</td>
</tr>
<tr>
<td>Bridging of fracture by callus or bone</td>
<td>10 (21)</td>
</tr>
<tr>
<td>or trabeculae</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Obliteration of fracture line and/or</td>
<td>3 (6)</td>
</tr>
<tr>
<td>cortical continuity</td>
<td></td>
</tr>
<tr>
<td>RUST score</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

| N, number; RUST, radiographic union scale in tibial fractures. |

measures across the various fracture types studied (Table 4). 
Based on the Quality Appraisal of Reliability Studies check- 
list applied to the latter 11 studies, 7 studies were considered 
to have an acceptable risk and 4 studies had an unclear risk of 
bias (see Appendix, Supplemental Digital Content 1, http:// 
links.lww.com/JOT/A820).

**TABLE 3. Results of 31 Interobserver Studies (Review B): 
Most Frequently Reported Outcome Measures of Radiologic 
Fracture Union**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Reported Outcomes, N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall union</td>
<td></td>
</tr>
<tr>
<td>Dichotomous</td>
<td>10</td>
</tr>
<tr>
<td>3-point scale</td>
<td>4</td>
</tr>
<tr>
<td>5-point scale</td>
<td>4</td>
</tr>
<tr>
<td>Specific radiographic criteria</td>
<td></td>
</tr>
<tr>
<td>Bridging of cortex, callus, or</td>
<td>12</td>
</tr>
<tr>
<td>trabeculae</td>
<td></td>
</tr>
<tr>
<td>Exact number of bridged cortices</td>
<td>3</td>
</tr>
<tr>
<td>Obliteration of fracture line</td>
<td>4</td>
</tr>
<tr>
<td>RUST/RUSS/RUSH scale</td>
<td>13</td>
</tr>
</tbody>
</table>

In total, the 31 interobserver studies tested the reliability of 73 fracture union outcome measures among observers. 

N, number; RUST, radiographic union scale in tibial fractures; RUSH, radiographic union score for hip fractures; RUSS, radius union scoring system.
that around one-third of the trials did not report fracture union criteria at all, limiting generalizability of their results.

To underline the importance of the reporting of observer characteristics and reliability statistics in RCTs, our second aim was to provide a general indication about observer differences that may exist between different specialist groups in rating fracture union on images. To this end, we performed a second review of interobserver studies and showed that interobserver reliability was often unsatisfactory between surgeons and radiologists who rated fracture union on images, across the various fracture types studied. In 3 orthopaedic trials,\textsuperscript{33–35} discussed in a previous review,\textsuperscript{1} radiologists gave more conservative measurements than surgeons for time to union and union rate. A possible explanation for such difference may be related to the knowledge of the patients’ clinical condition in trial setting; in one of the latter trials,\textsuperscript{34} in contrast to the other investigators, an independent radiologist had access to radiographs only and may have confirmed fracture union only when the signs of union were unambiguous on the images. Therefore, to interpret observer differences in trials, differences in the information available for each observer group, if any, should clearly be stated. In addition, in the literature, discrepancies between radiologists and surgeons are also observed in the radiologic classification of fracture types.\textsuperscript{36–38} A possible explanation for the latter difference may be related to the experience of surgeons with 3-dimensional views of different fracture patterns obtained during surgery.\textsuperscript{37} On the other hand, also within the same specialist group, large interobserver differences may exist, as shown in our overview (Table 4), which may be explained by differences in the level of training and experience and familiarity with the outcome scale used. In addition, the small

<table>
<thead>
<tr>
<th>Reference</th>
<th>Fracture Type</th>
<th>Cases (N)</th>
<th>Modality</th>
<th>Observers (N)</th>
<th>Outcome Scale</th>
<th>Test</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patel 2014–Series A</td>
<td>Distal radius (Cast)</td>
<td>35</td>
<td>X-ray</td>
<td>2 3</td>
<td>RUSS, score: 0–8</td>
<td>ICC</td>
<td>0.62 0.48 0.70</td>
</tr>
<tr>
<td>Patel 2014–Series B</td>
<td>Distal radius (ORIF)</td>
<td>35</td>
<td>X-ray</td>
<td>2 3</td>
<td>RUSS, score: 0–8</td>
<td>ICC</td>
<td>0.51 0.52 0.45</td>
</tr>
<tr>
<td>Hannemann 2015</td>
<td>Scaphoid waist</td>
<td>47</td>
<td>X-ray</td>
<td>2 2</td>
<td>Union, no, partial</td>
<td>κ</td>
<td>0.58</td>
</tr>
<tr>
<td>Hannemann 2013</td>
<td>Scaphoid waist</td>
<td>44</td>
<td>CT</td>
<td>1 2</td>
<td>Union, no, partial</td>
<td>κ</td>
<td>0.58</td>
</tr>
<tr>
<td>Dias 1988</td>
<td>Scaphoid waist</td>
<td>20</td>
<td>X-ray</td>
<td>4 4</td>
<td>Union, no, do not know</td>
<td>κ</td>
<td>0.39</td>
</tr>
<tr>
<td>Grewal 2013</td>
<td>Scaphoid</td>
<td>50</td>
<td>CT</td>
<td>1 1</td>
<td>Union, tenuous, partial</td>
<td>κ</td>
<td>0.80</td>
</tr>
<tr>
<td>Bhandari 2013 JOT*</td>
<td>Femoral neck</td>
<td>150</td>
<td>X-ray</td>
<td>3 3</td>
<td>Overall healing (Y/N)</td>
<td>ICC</td>
<td>0.22 0.17 0.21</td>
</tr>
<tr>
<td>Bhandari 2013 BMC*–Series A</td>
<td>Femoral neck</td>
<td>100</td>
<td>X-ray</td>
<td>3 3</td>
<td>Overall healing (Y/N)</td>
<td>ICC</td>
<td>0.60</td>
</tr>
<tr>
<td>Bhandari 2013 BMC–Series B</td>
<td>Intertrochanteric</td>
<td>100</td>
<td>X-ray</td>
<td>3 3</td>
<td>Overall healing (Y/N)</td>
<td>ICC</td>
<td>0.85</td>
</tr>
<tr>
<td>Chiavaras 2013</td>
<td>Intertrochanteric</td>
<td>150</td>
<td>X-ray</td>
<td>3 3</td>
<td>Overall healing (Y/N)</td>
<td>ICC</td>
<td>0.50</td>
</tr>
<tr>
<td>Franszone 2017</td>
<td>Tibia</td>
<td>30</td>
<td>X-ray</td>
<td>2 1</td>
<td>RUST, score: 4–16</td>
<td>ICC</td>
<td>0.92</td>
</tr>
<tr>
<td>Richards 2015</td>
<td>Tibia</td>
<td>36</td>
<td>X-ray</td>
<td>1 3</td>
<td>RUST, score: 4–12</td>
<td>κ</td>
<td>0.76</td>
</tr>
<tr>
<td>Bhattacharyya 2006</td>
<td>Tibia</td>
<td>35</td>
<td>CT</td>
<td>1 2</td>
<td>Overall healing (Y/N)</td>
<td>ICC</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Bold numbers are values > 0.7, indicating satisfactory agreement.

*Bhandari 2013 JOT used single radiographs at single time point to assess union, whereas Bhandari 2013 BMC used serial radiographs based on multiple time points.

N, number; RUSS, Radius union scoring system; RUST, radiographic union scoring system in tibial fractures; RUSH, radiographic union score in hip fractures; rad., radiologists; surg., surgeons; κ, kappa; Y, yes; N, no.
number of observers, variation in experience and training level, and the heterogeneity of applied outcome scales and fracture locations limit generalizability of all latter studies. Considering the above, the relative contribution of the observer specialty to observer difference in rating fracture union on images remains uncertain, and evidence is too limited to recommend one specialty over the other.

Our current approach has limitations. First, we focused on radiologic assessments, although clinical criteria are also often used to define fracture union. Nevertheless, radiologic union is an important end point in orthopaedic research, and its reliability should be examined. Second, study characteristics differed between included trials and interobserver studies (Table 1), which hampered aligning study results on specific fracture type level. Third, our first review was limited to the inclusion of only recently published RCTs because our aim was to make up-to-date recommendations on the reporting of observer characteristics in RCTs of surgical fracture treatment. This first review can therefore be seen as cross sectional, and a different period for article inclusion (eg, less recently published articles) may lead to different results. Finally, for our second review, we included only interobserver studies. Sometimes, other study designs such as case series report on measurements errors between observers too, although not reporting this in the abstract. Such studies remain undetected during literature screening because it is only possible to use search terms aimed at the title and abstract through Pubmed and Ovid.

Based on our 2 reviews, several recommendations can be made to reduce the risk of observer bias in orthopaedic trials on fracture healing treatment. First, we recommend trials to include at least 2 observers and to report observer characteristics and provide reliability statistics. Second, generalizability of interobserver studies is often low. Adherence to published guidelines for reporting reliability and agreement studies (GRASS) may improve generalizability and quality. Third, to reduce discrepancies between specialties, surgeons and radiologists should intensify collaboration and reach consensus on defining fracture union and nonunion. Some promising tools gain popularity to improve standardization of fracture healing assessment among specialists, including the radiographic union scale in tibial fractures (RUST), 24,29 radiographic union score for hip fractures (RUSH), 16,22 and radius union scoring system (RUSS). 28 Further studies remain required to evaluate their reliability and validity in general practice.

Blinding of observers of subjective outcomes reduces the risk of bias. 40 To this end, radiologic images can be digitally altered to mask the type of implant. 31 However, this may be rather labor intensive and may inadvertently mask parts of the fracture area affecting judgment. Alternatively, trials could integrate an independent panel of observers, including radiologist or surgeons who are not involved in other parts of the trial. Based on our review, most orthopaedic trials include only one observer, suggesting room for improvement.

In conclusion, discrepancies exist between radiologists and surgeons in the interpretation of radiologic findings during fracture healing. When conducting an orthopaedic trial of surgical fracture treatment, one should be aware that the specialty of the observer can contribute to bias, especially when only one observer is involved. Future research remains needed to evaluate the magnitude of the risk of bias associated with the observer specialty and specific level of training and experience.

REFERENCES
Invited Commentary

Interobserver Reliability in Imaging-Based Fracture Union Assessment—Two Systematic Reviews

In “Interobserver Reliability in Imaging-Based Fracture Union Assessment—Two Systematic Reviews,” the authors attempt to review the reporting of radiographic assessments of fractures in randomized comparative trials and in papers addressing the consistency in various scoring systems. Although the authors present a solid case for inclusion of the methodology of radiographic assessment in papers addressing radiographic union, the time frame is limited and the selection of studies allows for a wide variation in fracture locations, treatments, and radiographic modalities. This substantially limits the findings in the report. Fractures of metaphyseal and diaphyseal bone treated with completely rigid as well as flexible fixation that are evaluated with plain radiographs and computed tomography (CT) are all included. Despite these limitations, the conclusion that papers reporting on the radiographic outcomes of fractures should be transparent in their methodology is sound.

The inclusion of multiple locations and fixation types that lead to different patterns of healing makes any specific conclusions unusable. Diaphyseal injuries treated with intramedullary nailing have almost nothing in common with a metaphyseal fracture treated with rigid internal fixation, as it relates to assessing radiographic union. Similarly, the evaluation of a clinical nonunion by CT has little to do with the evaluation of an intertrochanteric hip fracture. Overall, these assessments are so disparate that no conclusion could possibly be drawn about future directions, and the heterogeneity of the studies truly precludes a cohesive combination of findings.1

With these substantial reservations in mind, the authors point out the following 2 relevant things: (1) Manuscripts with a radiographic assessment of union or healing score should include the specifics of evaluation, namely how many observers were used and their backgrounds, and (2) Clinicians who have vast experience with how bones heal may use that experience in their scoring of union. Readers of manuscripts should look to understand the methodology that leads to a diagnosis of “union” or