Evaluation of Sternal Bone Healing with Computed Tomography and a Quantitative Scoring Algorithm

Gregory S. Stacya,*, Osmanuddin Ahmeda, Arlene Richardsona, Brian M. Hatcherb, Heber MacMahonc and Jaishankar Raman

department of Radiology, University of Chicago Medical Center, Chicago, IL

director of Research, Biomet Microfixation, Jacksonville, FL

derpartment of Cardiothoracic and Vascular Surgery, Rush University Medical Center, Chicago, IL

Abstract: Objective: The exquisite bone detail offered by computed tomography makes it the ideal modality for evaluation of bone healing. However, few studies have investigated the normal computed tomographic appearance of the sternum after median sternotomy and, to the best of our knowledge, no computed tomographic classification of sternal healing has been proposed. Given the potential benefit of objective criteria, we propose a validated scoring classification of sternal healing using computed tomography for both clinical and investigational purposes.

Methods: Computed tomography scans from 20 patients who underwent a median sternotomy were evaluated for sternal healing at either 3 or 6 months postoperatively. Five anatomic locations along the sternum were selected using defined criteria, and a 6-point quantitative scale was developed to evaluate sternal healing. Independent radiologists read and scored each of the 5 locations on the sternum. Inter- and intra-observer variability was assessed by calculating the kappa statistics to measure the reliability of the scoring algorithm.

Results: Calculation of the kappa statistics indicated substantial agreement for intra-observer variability and substantial to almost perfect agreement for inter-observer variability. For intra-observer variability, the kappa statistics ranged from 0.591 to 0.802, and for inter-observer variability, the kappa statistics ranged from 0.590 to 0.969. When the two radiologists differed, the magnitude of the difference was no more than 1 or 2 points.

Conclusion: This simple system of evaluating sternal healing had high inter- and intra-observer reliability. Therefore, it may be considered a valid method for assessing sternal osteosynthesis for both clinical and investigational purposes.

Ultramini abstract: (49 words): Few studies have investigated the normal computed tomography appearance of the sternum after median sternotomy, and we knew of no computed tomography-based classification of sternal healing. Given the potential benefit of objective criteria, we designed and validated a scoring classification of sternal osteosynthesis for both clinical and investigational purposes.

Keywords: Computed tomography, sternum, wound closure, wound dehiscence, wound healing.

INTRODUCTION

Median sternotomy is the most commonly performed osteotomy worldwide [1] and is used primarily for cardiac surgery. This osteotomy is typically closed at the completion of the procedure with wire cerclage, although in recent years, there has been increasing interest in rigid plate fixation of the sternum [2-10]. The assessment of sternal healing has been limited by crude methods of clinical assessment rather than a truly objective test—unless there was evidence of sternal dehiscence. Computed tomography (CT) scans are used in the setting of sternal wound complications to assess dehiscence, presence of infection, and the need for sternal re-exploration [11, 12]. The exquisite bone detail offered by CT makes it the ideal modality for evaluation of bone healing. Few studies, however, have investigated the normal CT appearance of the sternum after median sternotomy and, to the best of our knowledge, no CT classification of sternal healing has yet been proposed. Given the potential benefit of objective criteria, a validated scoring classification of sternal healing using CT for both clinical and investigational purposes is proposed. The objective of this study was to develop a simple scoring algorithm for evaluating sternal bone healing using CT scans and to determine the repeatability and reliability of this algorithm.

PATIENTS AND METHODS

Patient Population

CT scans from patients enrolled in a prospective, randomized multicenter clinical trial (Clinicaltrials.gov NCT00819286) were used to develop a quantitative
algorithm for evaluating sternal bone healing [5, 7]. Following patient consent and Institutional Review Board (IRB) approval, patients underwent a median sternotomy as a standard component of a cardiac surgical procedure. These were patients considered at increased risk for postoperative sternal complications. Sternal closure was performed using either rigid plate fixation or wire cerclage in the setting of a randomized, prospective study. IRB approval was obtained for the current study, which involved review of CT scans obtained 3 or 6 months after surgery.

CT Scan Parameters

The original CT scans consisted of transverse images acquired from the lung apices to the costophrenic sulci, with thin (3 mm or less) contiguous transverse slices obtained through the entire sternum. All scans were obtained with suspended respiration, when possible. The Gantry tilt technique was not used. Images were archived in DICOM format and submitted to a core laboratory for scoring. Bone windows (window level 300 HU, window width 1,500 HU) were used for selection and scoring of individual slices.

Development of Scoring System

Using the CT examinations of 11 patients, two radiologists (G.S.S., O.A.) reviewed the various patterns of sternal healing depicted on the scans. Features analyzed included sternal approximation and alignment, gap mineralization between the sternal halves, and imaging characteristics of the osteotomy margins (ie, absence or presence of sclerosis, resorption, or contour irregularity). Using these basic imaging features, a simple 6-point scale was developed and applied to analyze the sternal healing of 20 additional patients.

Analysis of Sternal Healing

CT scans of 20 additional patients with various degrees of sternal healing were selected for analysis. Before distribution of the scans to radiologists (G.S.S., A.R.) for interpretation, a radiologist (O.A.), who was not otherwise involved with scoring, selected 5 anatomic levels to allow evaluation of healing along the entire length of the sternum: Level 1 (Manubrium at the first sternocostal articulation, using the first slice that does not show the clavicles [scrolling from cranial to caudal]); Level 2 (Upper end of sternal body, using first slice below the second costal cartilage that does not show the sternomanubrial junction [scrolling from cranial to caudal]); Level 3 (Mid-upper sternal body, using the slice through the lowest margin of the third costal cartilage); Level 4 (Mid-lower sternal body, using the slice through the lowest margin of the fourth costal cartilage); and Level 5 (Lower end of sternal body, using the first slice at the level of the fifth costal cartilage that does not show the xiphoid process or xiphoid-body junction [scrolling from caudal to cranial]).

If the presence of a streak artifact from the wires or sternal plates at the original level would prohibit accurate scoring, a level within two slices of the original level could be selected for scoring. Fig. (1) is a representative image of
the slices selected for scoring. After the levels were selected and designated for scoring, the 20 complete CT scans were provided separately to two radiologists for review. Both radiologists were blinded to the other’s responses. Each radiologist graded the selected levels independently using the previously mentioned 6-point quantitative scale (0-5). Both 3- and 6-month scans were evaluated to assess the progression of sternal healing. Ten patients from each timepoint were included in the analysis.

### Determination of Inter- and Intra-Observer Variability

Inter-observer variability was assessed by having separate radiologists independently score CT scans from the same patients. Intra-observer reliability was assessed by representing the cases to the same radiologists, although in different order, during a separate interpretation session. Patient identifying data were removed from the CT images before interpretation.

Kappa statistics were calculated to measure inter- and intra-observer variability [13]. Kappa statistics are commonly used to estimate observer agreement and can be interpreted as the proportion of agreement among raters after chance agreement has been removed [14]. In order to distinguish degrees of agreement, the weighted form of the kappa statistic was used. In this way, the analysis took into account disagreements that were, for example, just one category—rather than several—away.

### RESULTS

#### Scoring System

The results of scoring system development are shown in Table 1. The 6-point quantitative scale consisted of the following categories: (0) features of nonunion, (1) no definite healing or indeterminate features, (2) features of minimal or early healing, (3) features of mild healing, (4) features of moderate healing, and (5) features of complete healing. Scores 1-4 also consisted of subscores, depending on additional features described in Table 1, but these subscores were not considered for purposes of statistical analysis. A score of 5 represented complete bony union, with the cross-section of the sternum appearing essentially as normal bone (ie, sternal halves well aligned, with no perceptible gap or perceptible osteotomy margin or only minimal residua of previous osteotomy). A score of 4 represented healing with bridging bone along 50% or more of the anteroposterior dimension of the sternal halves. A score of 3 represented healing with bridging bone along less than 50% of the anteroposterior dimension of the sternal halves. A score of 2 represented minimal but convincing features of healing, with a thin bridge of bone connecting the sternal halves anteriorly or posteriorly or faint, poorly-defined mineralization between noncontacting or nearly contacting sternal halves. A score of 1 represented no contact between sternal halves or gap mineralization but nonsclerotic, irregular, or concave osteotomy margins. A score of 0 implied nonunion, with no contact or mineralization between the sternal halves, but osteotomy margins were nonsclerotic (1A), concave (1B), or irregular (1C).

### Table 1. Six-point scale used to evaluate sternal bone healing.

<table>
<thead>
<tr>
<th>Score</th>
<th>Definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No sign of healing: nonunion</td>
<td>No contact between sternal halves, absence of gap mineralization, and sclerotic osteotomy margins similar to that of cortical bone</td>
</tr>
<tr>
<td>1</td>
<td>Indeterminate</td>
<td>No contact or mineralization between the sternal halves, but osteotomy margins were nonsclerotic (1A), concave (1B), or irregular (1C)</td>
</tr>
<tr>
<td>2</td>
<td>Signs suggesting minimal or early healing</td>
<td>Faint mineralization between noncontacting sternal halves (2A), a thin (1 mm) bridge of bone connecting the sternal halves anteriorly or posteriorly (2B), or near bone-on-bone contact between the sternal halves, with sclerotic osteotomy margins (2C)</td>
</tr>
<tr>
<td>3</td>
<td>Mild synthesis</td>
<td>Bridging bone (ie, no perceptible gap) along less than 50% of the anteroposterior dimension of the sternal halves, with the sternal halves either offset in the anteroposterior dimension (3A) or aligned in the anteroposterior dimension (3B)</td>
</tr>
<tr>
<td>4</td>
<td>Moderate synthesis</td>
<td>Bridging bone (ie, no perceptible gap) along 50% or more of the anteroposterior dimension of the sternal halves, with the sternal halves either offset in the anteroposterior dimension (4A) or aligned in the anteroposterior dimension (4B), with visible remnants of the previous osteotomy</td>
</tr>
<tr>
<td>5</td>
<td>Complete synthesis</td>
<td>Sternal halves were well aligned, and the appearance was that of essentially normal bone, without a gap or visible osteotomy margin</td>
</tr>
</tbody>
</table>

Sternal Healing

Comparison of 3- and 6-month CT scan scores demonstrated the sensitivity of the methods in measuring sternal bone healing over time. At 3 months, the mean (SD) CT score was 1.9 ± 0.8, which is representative of minimal or early healing. At 6 months, the mean (SD) CT score was 3.3 ± 0.9 (P<0.001), which is characteristic of mild synthesis. It is also important to note that none of these patients had clinical signs of sternal wound problems.
weighted Kappa statistics were calculated separately for each site (5 total) and for each of two reads. Calculation of the kappa statistics indicated moderate to substantial agreement for inter-observer variability (Table 2) and moderate to almost perfect agreement for intra-observer variability (Table 3). For inter-observer variability, the kappa statistics ranged from 0.591 to 0.802 and, for intra-observer variability, they ranged from 0.590 to 0.969. When the readers differed, the magnitude of the difference was no more than 1 or 2 points. Intra-observer agreement for Readers 1 and 2 was 59% and 68% (no difference in scores); intra-observer disagreement was 35% and 24% (1-point difference in scores), and 6% and 8% (2-point difference in scores). Inter-observer agreement was 87% and 68% for Reads 1 and 2 (no difference in scores); inter-observer disagreement was 9% and 24% (1-point difference in scores), and 4% and 8% (2-point difference in scores).

Comment

Sternotomy is the most commonly performed osteotomy worldwide (>1 million annually) [15], and it is important to have an objective, standard method of assessing healing after these procedures. Unlike other specialties, in which radiographic standards have been developed to assess bone healing, no standards exist to assess bony healing after a sternotomy [16-19]. The methods developed for other orthopaedic specialties have clinical implications, and may be used to determine when patients may return to normal activities or when hardware may be removed.

Indeed, most evaluations are clinical, and CT scans are used only when wound problems supervene. Confirmation of bone healing in all of these applications is of value, as nonunion can lead to complications such as infection, pain, and limited function [20-24]. CT is an excellent imaging modality for evaluation of bone healing. Modern multi-channel scanners allow thin slices and minimization of metal artifact compared with earlier-generation scanners. CT has been used to assess healing of fractures in the pre- and post-operative settings and fracture complications [25-29]. We have proposed a simple, validated system of scoring sternal osteosynthesis after sternotomy that can be used by radiologists, surgeons, and other clinicians and investigators for both clinical and research purposes, with kappa statistics indicating moderate to almost perfect agreement. The sternum can have a variety of appearances on CT following sternotomy and fixation. This variety necessitates a scoring system that is comprehensive, but simple in its methodology, essentially evaluating various degrees of (1) contact and alignment between the sternal haves, (2) gap mineralization, and (3) sclerosis of the osteotomy margins. By evaluating these 3 simple variables and assigning a score ranging from nonunion (0) to complete healing (5), one can simultaneously achieve a reproducible estimate of healing or lack of healing while maintaining the granularity needed to compare healing over short time intervals.

Bone healing after fracture or osteotomy is a complicated process involving osteoprogenitor cells and bone formation stimulated by local factors, resulting in endosteal and periosteal callus and, ultimately, complete fusion of bone parts. Careful alignment of the bone and maintenance of that bony apposition is critical for healing. Bony movement and separation of as little as 2 mm can result in a critical-sized

**Fig. (2).** Illustrations and axial computed tomographic scans depicting the 6-point scale and corresponding subscales used to evaluate sternal bone healing.

**Inter- and Intra-Observer Variability**

Weighted Kappa statistics were calculated separately for each site (5 total) and for each of two reads. Calculation of the kappa statistics indicated moderate to substantial
Table 2. Inter-observer kappa statistics. Kappa scores indicate degree of agreement between the two readers: slight (0.01 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80), or almost perfect (0.81 to 0.99). Each anatomic level was evaluated twice by each reader, and two kappa statistics for each read are reported.

<table>
<thead>
<tr>
<th>Anatomic Level</th>
<th>First or Second Read</th>
<th>Weighted kappa</th>
<th>95% Lower Confidence Limit</th>
<th>95% Upper Confidence Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.698</td>
<td>0.537</td>
<td>0.859</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.745</td>
<td>0.555</td>
<td>0.936</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.591</td>
<td>0.413</td>
<td>0.770</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.712</td>
<td>0.506</td>
<td>0.917</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.674</td>
<td>0.522</td>
<td>0.827</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.675</td>
<td>0.522</td>
<td>0.828</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.802</td>
<td>0.683</td>
<td>0.921</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.779</td>
<td>0.612</td>
<td>0.945</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.681</td>
<td>0.473</td>
<td>0.888</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.795</td>
<td>0.635</td>
<td>0.956</td>
</tr>
</tbody>
</table>

Table 3. Intra-observer kappa statistics. Kappa scores indicate degree of agreement for a single reader: slight (0.01 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80), or almost perfect (0.81 to 0.99). Each anatomic level was evaluated twice by each reader, and two kappa statistics for each read are reported.

<table>
<thead>
<tr>
<th>Anatomic Level</th>
<th>First or Second Reader</th>
<th>Weighted Kappa</th>
<th>95% Lower Confidence Limit</th>
<th>95% Upper Confidence Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.764</td>
<td>0.578</td>
<td>0.951</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.816</td>
<td>0.635</td>
<td>0.996</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.913</td>
<td>0.826</td>
<td>0.999</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.590</td>
<td>0.351</td>
<td>0.829</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.880</td>
<td>0.753</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.777</td>
<td>0.621</td>
<td>0.933</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.969</td>
<td>0.910</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.714</td>
<td>0.569</td>
<td>0.859</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.968</td>
<td>0.906</td>
<td>1.000</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.605</td>
<td>0.378</td>
<td>0.831</td>
</tr>
</tbody>
</table>

gap and nonunion [30, 31]. Studies have shown greater chest pain in patients with sternal nonunion compared to patients with sternal healing. Sternal separation in more than one location along the sternum also resulted in greater pain [32]. Many patients undergoing a median sternotomy are at risk for sternal nonunion, which can be affected by the degree of stability of fixation, reduction of blood supply (eg, as seen after internal mammary artery harvest), and comorbidities such as diabetes and smoking [33-35]. The use of immuno-suppressive drugs (eg, steroids) and the presence of osteoporosis also affect bone healing. Although it is well known that clinical union of a fracture precedes union on imaging studies, we propose that CT can estimate sternal healing based on the presence or absence of a radiolucent gap or visible mineralization between sternal halves and the appearance of the osteotomy margins.

As with other bones, a persistent gap between sternal halves with well-defined sclerotic osteotomy margins is presumed to represent nonunion (score of 0 in our study), whereas the appearance of normal or minimally deformed bone without a perceptible osteotomy margin is presumed to represent healed bone (score of 5 in our study). A score of 1 represents an indeterminate stage without conclusive signs of healing or nonunion, and scores of 2 through 4 show progressively advanced stages of healing. While a score of 2 represented scans with evidence of bone formation between sternal halves, the amount of interposed bone was limited to
such a degree that a classification of union by imaging should be reserved for scores of 3 and higher. Using this threshold for sternal union, only 2 of the 40 scans evaluated had inter- or intra-observer variability that resulted in disparate classifications of sternal union and nonunion for the same patient. Thus, reporting both the mean scores and union rates provides valuable information; both should be considered when presenting results using this method. The incremental differences in each of the scores on the 6-point scale were slight, and 1-point differences in score assignment occurred. Two-point differences were uncommon. Of note, there were no disagreements of more than 2 points.

This study is an important step in redefining the understanding of sternal healing after sternotomy. In addition, this is an attempt to develop a simple standard that can be used universally. Bone healing is a desirable but elusive hard end point with far-reaching clinical implications. For instance, patients after sternotomy are often considered as an arbitrary length of time under the assumption that the sternum heals at 8 weeks. This study gives us a more realistic timeline of sternal healing, which might result in re-evaluation of the advice routinely given to heart surgery patients. Avoidance of complications in predisposed patients may be guided by the findings of this study as well.

Our study has several limitations. First, although the scores assigned to the various CT appearances of the stages of sternal healing seem logical to us from a physiological standpoint, we do not have a histologic, biomechanical, or clinical “gold standard” with which we can compare the images. As stated earlier, clinical union typically precedes union depicted on imaging studies; therefore, CT may underestimate the degree of sternal union. The presence of small gaps may not be clinically significant [34]. Furthermore, it is conceivable that two united sternal halves could be so perfectly apposed to one another as to mimic complete union; however, based on our experience, we think this is unlikely.

Second, although CT scans obtained at both 3 and 6 months after sternotomy were evaluated, longitudinal imaging studies of the same patients were not available; therefore, we could not assess progressive healing (or lack thereof) in any individual patient. However, our data showing that the scores for patients with CT scans obtained 6 months after surgery were significantly higher than those of scans obtained 3 months after surgery support our methodology. Third, the CT scans used for analysis were originally part of a separate multicenter clinical trial, which introduced variability into the scans based on slightly different technical parameters between institutions; however, as this variability mimics daily practice at different institutions, we do not feel that this significantly biased or otherwise contaminated our data or interpretations. Furthermore, the goal of this investigation was to create a simple system that can be used universally without regard to medical subspecialty or specific scanner type. Fourth, although we made every attempt to standardize the levels at which sternal healing was scored, the presence of metal artifact occasionally necessitated a slight adjustment of the anatomic level to be scored; however, because the levels were predetermined by an independent radiologist not involved with scoring, we think that the adjustment did not affect the kappa statistics.

In conclusion, our results show that our system of scoring for sternal healing has high inter- and intra-observer reliability and, therefore, may be considered a valid method for assessing sternal healing for both clinical and investigative purposes.

CONFLICT OF INTEREST

The study was funded by Biomet Microfixation. All funds used to support the study, including the CT scans and their reporting, were included in this. Dr Raman is an advisor to Biomet Microfixation, and Dr Hatcher is an employee of Biomet. The authors had full control of the design of the study, methods used, outcome parameters and results, analysis of data, and production of the written report.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES


