Reliability of the Penny and Beit CURE Radiologic Classifications of Pediatric Patients with Chronic Hematogenous Osteomyelitis in the Philippine General Hospital

Karla Teresa S. Araneta, MD[1] and Juanito S. Javier, MD, MChOrth[2]

[1]*Department of Orthopedics, Philippine General Hospital, University of the Philippines Manila*
[2]*Section of Pediatric Orthopedics, Department of Orthopedics, Philippine General Hospital, University of the Philippines Manila*

ABSTRACT

Objective. This study aimed to evaluate the inter- and intraobserver reliability of the Penny and Beit CURE radiologic classifications of pediatric patients with Chronic Hematogenous Osteomyelitis (CHOM) in the Philippine General Hospital (PGH).

Methods. Thirty-four pre-operative radiographs of PGH pediatric patients with CHOM were classified by seven orthopedic surgeons using both Penny and Beit CURE Classification systems. Two sets of radiographs were sent to the surgeons twice, four weeks apart, to classify. The Fleiss and Cohen κ statistics were used to determine inter- and intraobserver reliabilities, respectively.

Results. The Penny Classification had a slight to fair interobserver reliability (Fleiss κ = 0.17 and 0.24) and a fair intraobserver reliability (Cohen κ =0.35) with a 49.58% average intraobserver agreement. The interobserver reliability when including all Beit CURE classification subtypes was fair (κ = 0.28 and 0.31). This improved to moderate (κ = 0.41 and 0.54) when using only the four main types of the Beit CURE classification with a 77.31% intraobserver agreement.

Conclusion. The Beit CURE classification for pediatric CHOM had higher inter- and intraobserver agreement rates than the Penny classification. Further improvement in reliability can be made by combining B2 and B3 subtypes under the Beit CURE classification.

Key Words: chronic hematogenous osteomyelitis, Beit CURE, Penny classification

INTRODUCTION

Chronic osteomyelitis (COM) represents a major health problem in the developing world due to its significant morbidity and low mortality rate. It places a substantial burden on developing countries' health services, leading to hospitalization and prolonged antibiotic administration and, sometimes, causing permanent disability.[1,2] Chronic hematogenous osteomyelitis, an infection in bone originating from bacteremia or septicemia, and lasting for three or more months, have more consolidated data in the medical literature and is considered a predominantly pediatric disease with 85% of patients aged below 17 years.[3]

Surgery has been the mainstay of treatment for this disease. Initial pre-operative planning requires establishing the correct diagnosis using a radiologic classification system that defines the natural history, which assists in planning surgical treatment.[4]

3rd place in Research Paper in the Philippine Orthopedic Association (POA) 68th Annual Congress Residents Research Forum on November 15-18, 2017, in EDSA Shangri-La, Mandaluyong City, Philippines.

Corresponding author: Karla Teresa S. Araneta, MD
Department of Orthopedics
Philippine General Hospital
University of the Philippines Manila
Taft Avenue, Manila 1000, Philippines
Email: ksaraneta1@up.edu.ph

Table 1. Penny and Beit CURE Radiologic Classifications for Pediatric Osteomyelitis

Classification		Description	Treatment
Penny	I. Typical	Sequestrum and involucrum	Remove sequestrum
	II. Atrophic	No involucrum (periosteal death)	Observe for 3-6 months; if no involucrum, then grafting or transport
	III. Sclerotic	Fusiform, dense cortical thickening, the medullary canal may be obliterated; sequestrum may be hidden	Search for sequestrate, may be difficult to remove cortical window
	IV. Cortical	Localized sequestrum in the cortex	Remove sequestrum
	V. Walled-off abscesses	Partial resorption of sequestrate, leaving well-defined lucencies in involucrum	Saucerization, look for sequestrate
	VI. Multiple microabscesses	Similar to walled-off abscesses, small and numerous	Difficult to remove, consider antibiotics and observation
	VII. Metaphyseal	Single or multiple abscesses with sclerotic margin	Saucerization and curettage
Beit CURE	A. Brodie's type	No sclerosis, one or more abscesses	Drilling and curettage of abscess OR conservative treatment with six weeks of antibiotics
	B. Sequestrum and Involucrum (B1-4)	B1. Localized cortical sequestrum	Sequestrectomy and curettage
		B2. Sequestrum with normal involucrum	
		B3. Sequestrum with sclerotic involucrum	
		B4. Sequestrum without structural involucrum	Sequestrectomy and curettage AND stabilization with plaster, traction, or external fixator
	C. Sclerotic	Extensive sclerosis and abscesses	Drainage and curettage of any collection and long-term antibiotics (6-week minimum)
	Unclassifiable		
	Physeal damage: add P if proximal physis or D if distal physis is damaged		

Unfortunately, there has been no universally accepted classification system, although more than ten classification systems have been developed in the last four decades. Shortcomings have included problematic host stratification, failure to keep up with current treatment options, lack of clinical applicability, or poor inter-observer reliability.[5,6]

In the last decade, two classification systems have shown to be specifically useful in classifying pediatric patients with chronic osteomyelitis: 1) Penny Classification; and 2) Beit CURE Classification.[1,7] The characteristics and treatment stratification for both classifications is summarized in Table 1. The pathoanatomic features in both the Penny and Beit CURE classifications of chronic osteomyelitis help the surgeon judge the stability of the involved bony segment and determine the timing and extent of the proposed surgery to predict possible complications and/or the projected need for further procedures.[4] Both classifications describe a spectrum of radiographic findings related to the 1) extent of bony necrosis and 2) host's response.[4] While there is significant overlap between these, the Beit CURE classification also includes assessing physeal damage at the proximal and/or distal end of the bone and a category for cases that cannot be classified based on their scheme.[8-10]

In terms of inter- and intraobserver variability, the measure of reliability is essential to any injury classification system.[11,12] No studies have yet been done assessing the reliability of the Penny Classification, much less even compared the two classification systems in terms of observer variabilities.

This study aimed to evaluate the inter- and intraobserver reliability of the Penny and Beit CURE radiologic classifications of PGH pediatric patients with chronic hematogenous osteomyelitis. No reliability studies have yet been done comparing the Penny and Beit CURE radiologic classifications. This is the first study to date determining the reliability of both classification systems. This study's results could significantly impact which classification system we adopt in our setting, influencing the management of this extraordinarily complex set of patients. To date, there is still no local institutional, let alone national clinical practice guideline on classifying CHOM. If validated, the more reliable classification system could be used as a systematic guide to help our local orthopedic surgeons in pre-operative planning. This is a stepping-stone to finally developing a systematic and outcome-based approach in treating chronic hematogenous osteomyelitis in children, something that has not yet been achieved in any institution.

MATERIALS AND METHODS

Inclusion criteria

A list of patients aged 17 and younger admitted at PGH from 2012-2016 with the diagnosis of chronic hematogenous osteomyelitis was made based on the medical information system (MIS) of the Orthopedic Department and the medical records section (ICD 10 codes – K10.2, M46.2, M86.0-M86.9, M90.2). Their pre-operative radiographs (AP and lateral views) were then gathered from the PGH

Radiology Section. Additional radiographs were also collected from non-service cases done in PGH by the senior author. In all cases, the patients had been affected by osteomyelitis for at least six months.

Exclusion criteria

Radiographs of acute, subacute, and recurrent multifocal OM and septic arthritis were not included in this study. Post-operative radiographs of patients with COM were not included. Finally, radiographs of patients with concurrent radiographic findings of the affected bone, such as congenital osseous anomalies, tumors, pathologic fractures, etc., were not included.

Study procedure

The radiographs of the selected cases were photographed using a high-resolution 18-megapixel camera and compiled (as a PDF file) in a USB together with an instructional sheet, consent forms, two sets of answer sheets (1 answer sheet for the Beit CURE classification and one answer sheet for the Penny Classification) and a case example to clarify the methodology. The USB also included a file with a reference guide for both classification systems if the raters feel the need to review the classifications.

The materials were placed in an envelope and sent to 7 Orthopedic Surgeons in PGH (2 pediatric orthopedic consultants and 5 Senior residents who have rotated in the Ortho-Pedia Service). Each observer was tasked to stage the affected bone using both classifications.

Once completed, the data was sent to an independent authorized person for filing. The same materials were sent again after four weeks with the radiographs in a different order to determine the intra-observer reliability.

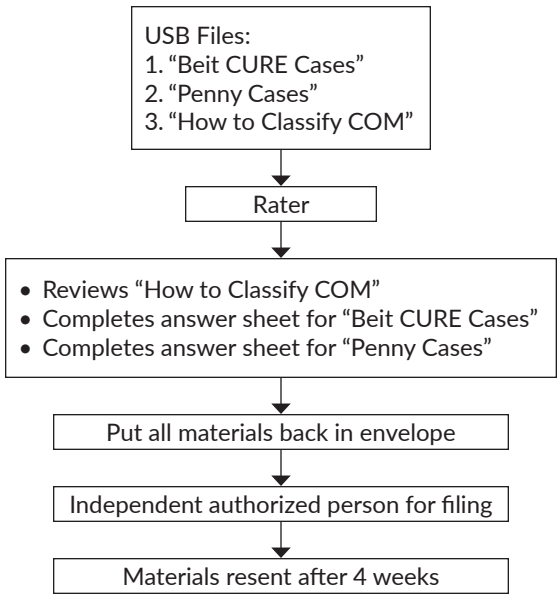

Figure 1. Flowchart of Methodology.

Sample size

All radiographs from 2012-2016, fitting the inclusion criteria, as stated above, were included in the study. The researchers further assessed cases to ensure that each category was included and that a fair spectrum of radiographs included typical cases to those that are more difficult to classify.

Statistical Analysis

The Fleiss κ statistic was used for measuring inter-observer reliability of both classification systems, while for intraobserver reliability, the Cohen κ was used. To maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels were assigned to the corresponding kappa ranges, as shown in Table 2.

Table 2. Percentage of agreement at a variety of κ statistics levels (Landis and Koch)[12]

κ Value	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.80-1.00	Almost Perfect

Ethical Considerations

The study protocol was approved by the University of the Philippines Manila Research Ethics Board (UPMRB) Panel. All patient information was kept anonymous and confidential. Also, the digital photographs were edited to conceal any patient identifying markers on the radiographs.

The primary investigator provided funding for the research. There were no other sources of financing. There were no conflicts of interest in the conduct of this study.

Consent was obtained from each orthopedic surgeon upon receiving the folder containing the soft copy of the radiographs, instruction sheet, and answer sheets.

RESULTS

Radiographs

A total of thirty-four radiographs of pediatric chronic hematogenous osteomyelitis were used in this study. All subtypes under the Penny and Beit CURE, as evaluated by the authors, were represented. Most studies support the use of approximately this number of raters and samples due to sufficient variability without inducing fatigue and loss of attention among the raters.[1]

Interobserver Reliability

The summary of Fleiss κ statistics for both classifications is found in Table 3. The interobserver reliability of the Penny Classification has a Fleiss κ statistics of 0.24 for the

1st batch and 0.17 for the 2nd batch. According to Landis and Koch κ statistical levels (Table 2), these have only fair and slight strength of agreement, respectively.

The Beit CURE classification's interobserver reliability was analyzed first using the main types (A, B, C, and Unclassified) and then analyzed further by including the subtypes under B (A, B1, B2, B3, B4, C, and Unclassified). When all the Beit CURE Classification's subtypes were included, Fleiss κ statistics of 0.28 and 0.31 were obtained. This only has a fair strength of agreement. However, when analyzing only the main types under Beit CURE Classification, the Fleiss κ statistics improved to 0.41 and 0.54, both with a moderate strength of agreement.

Intraobserver Reliability

Intraobserver reliabilities for both classifications are summarized in Tables 4, 5, and 6. The intraobserver reliability of the Penny Classification had a Cohen κ of 0.35. Intraobserver agreement averaged at 49.58% between the 1st and 2nd set. Both the main and subtypes of the Beit CURE Classification had moderate intraobserver reliability (κ = 0.51 and κ = 0.41, respectively).

Table 3. Results of the interobserver reliability for Penny and Beit CURE classification

Classification	Set 1		Set 2	
	κ	Strength	κ	Strength
Penny	0.24	**Fair**	0.17	**Slight**
Beit CURE (main Types)	0.41	**Moderate**	0.54	**Moderate**
Beit CURE (with subtypes)	0.28	**Fair**	0.31	**Fair**

Table 4. Results of the intraobserver reliability of the Penny classification

Rater	Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
1	50.00	23.27	0.3484	0.0807	4.32	0.0000
2	52.94	22.23	0.3949	0.0839	4.70	0.0000
3	35.29	16.70	0.2233	0.0726	3.07	0.0011
4	50.00	29.67	0.2891	0.0872	3.31	0.0005
5	41.18	17.39	0.2880	0.0754	3.82	0.0001
6	58.82	18.60	0.4942	0.0768	6.43	0.0000
7	58.82	28.55	0.4237	0.0858	4.94	0.0000
Average	**49.58**	**22.34**	**0.3500**		Fair	

Table 5. Results of the intraobserver reliability of the Beit CURE classification (main types)

Rater	Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
1	67.65	37.11	0.4856	0.1023	4.75	0.0000
2	85.29	0.5981	0.1345	0.1345	4.45	0.0000
3	70.59	49.05	0.4228	0.1086	3.89	0.0000
4	82.35	49.22	0.6525	0.1112	5.87	0.0000
5	70.59	36.85	0.5342	0.0942	5.67	0.0000
6	85.29	41.61	0.7481	0.1085	6.89	0.0000
7	79.41	49.83	0.5897	0.1164	5.07	0.0000
Average	**77.31**	**37.75**	**0.5100**		Moderate	

Table 6. Results of the intraobserver reliability of the Beit CURE classification (with subtypes)

Rater	Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
1	26.47	12.02	0.1600	0.0596	2.76	0.0029
2	61.76	22.49	0.5067	0.0862	5.88	0.0000
3	38.24	19.55	0.2323	0.0790	2.94	0.0016
4	61.76	23.18	0.5023	0.0761	6.60	0.0000
5	47.06	14.53	0.3806	0.0674	5.65	0.0000
6	70.59	16.18	0.6491	0.0718	9.05	0.0000
7	55.88	18.43	0.4592	0.0773	5.94	0.0000
Average	**51.68**	**18.05**	**0.4100**		Moderate	

Table 7. Results of the interobserver reliability for classification of physeal damage under Beit CURE classification

	Set 1		Set 2	
Classification	κ	Strength	κ	Strength
Physeal Damage	0.40	Fair	0.38	Fair

Table 8. Results of the intraobserver reliability for classification of physeal damage under Beit CURE classification

Rater	Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
1	73.53	51.82	0.4506	0.1228	3.67	0.0001
2	85.29	76.56	0.3727	0.1511	2.47	0.0068
3	91.18	44.29	0.8416	0.1264	6.66	0.0000
4	73.53	34.95	0.5931	0.1190	4.98	0.0000
5	85.29	35.99	0.7703	0.1222	6.30	0.0000
6	73.53	47.58	0.4950	0.1136	4.36	0.0000
7	82.35	62.63	0.5278	0.1243	4.24	0.0000
Average	**80.67**	**50.55**	**0.5800**		Moderate	

Assessment of physeal damage was analyzed separately from the Beit CURE main and subtypes. Analyzed alone, interobserver reliability of Physeal Damage had a Fleiss κ statistics of 0.40 or only fair strength of agreement among raters. Its intraobserver reliability was measured with a Cohen κ statistic of 0.58 (moderate strength) with an average intraobserver agreement of 80% between the 1st and 2nd set.

DISCUSSION

Using Fleiss κ statistic, the interobserver reliability when including all subtypes of the Beit CURE classification was measured at 0.28 and 0.31, indicating only fair agreement among surgeons. If we condensed this classification into its four main types, interobserver reliability improved to moderate (κ = 0.41 and 0.54). This indicates the difficulties associated with agreeing to a particular B subtype of the Beit CURE classification. Jones et al., in their study in 2009, already observed that combining B2 and B3 subtypes would improve the interobserver reliability since different observers may have a lower threshold than others for calling an involucrum sclerotic and because of the lack of a specific limit between a normal and sclerotic involucrum. When we performed an additional analysis combining B2 and B3 subtypes, Fleiss κ slightly improved to 0.32 (moderate). We further combined B1, B2, and B3 subtypes and analyzed the interobserver reliability resulting in a similar Fleiss κ statistic of 0.32. It did not differ on the interobserver agreement whether we combined B2 and B3, B1, B2, and B3. This suggests that raters could distinguish B1 from the other B subtypes and that we would have to combine all B-subtypes with improving the strength of the interobserver agreement substantially. To better modify this system, we can combine B2 and B3 separate from the B1 and B4 subtypes since adding B1 to the combination did nothing to improve the reliability. Regarding treatment, B2 and B3 would most likely have the same surgical strategy of creating a bone window or trough in addition to a sequestrectomy. Improving this classification's reliability based on these modifications would provide a less cumbersome way of selecting treatment options and assessing clinical outcomes.

Our overall inter- and intraobserver reliability was lower than that obtained by Jones et al.[1] wherein they found substantial interobserver agreement using the main types (κ = 0.73 ad 0.84), moderate interobserver agreement when using all subtypes (κ = 0.53 and 0.54) of the classification and a substantial intraobserver agreement of 0.86. This could be primarily because we included an Unclassified type under the Beit CURE classification to account for radiographs that raters were not confident in classifying as sclerotic type or sequestrum-involucrum type. Also, Jones' study had fewer observers in their pool and more years in orthopedic experience than ours.

Assessment of physeal damage was done separately since a previous study has already shown that combining this with the Beit CURE subtypes gave poor inter and intra-observer reliability. In this study, the intraobserver reliability of physeal damage assessment ranged from 73 to 91%, with an average of 81% agreement. The results of this sub-analysis were similar to that of Jones et al.[1] However, our interobserver reliability was only fair (κ = 0.40) compared to the substantial agreement found in their paper.[1] Orthopedic consultants can more readily and accurately assess physeal damage than orthopedic residents, who could again explain this difference.

The Penny Classification evaluation revealed only a slight to fair interobserver agreement (κ = 0.17 and 0.24). Its intraobserver reliability was also fair (Cohen κ = 0.35) with a 49.58% average intraobserver agreement.

The problem with the Penny Classification as reflected in only the slight agreement between surgeons is its overlapping definitions per subtype such that, for example, in Figure 2, a cortical (type IV) bone can also present with an involucrum

Figure 2. Sample radiograph identified by three observers as Penny Type I and three other observers as Type IV.

Figure 3. Sample radiograph identified as Penny type I, III, and V by different observers.

making it challenging to differentiate from a typical sequestrum-involucrum (type I). As illustrated in Figure 3, another confusion may be walled off abscesses (V), appearing to present with a dense cortical thickening (type III).

The Beit CURE classification has somewhat addressed these significant overlapping pathoanatomic differences in the Penny classification. They specifically create a subtype for the typical sequestrum-involucrum types seen in a Penny classification and then follow a step-wise approach to eliminating other features on the radiograph. To illustrate, in the Beit CURE classification, the lack of a sequestrum seen on the radiograph is automatically referred to as either Brodie's type or Sclerotic type. This is not implicit in the Penny Classification. The same radiograph in Figure 2 was identified more consistently in the Beit CURE classification as B2 (by five observers) and B3 (by only two observers). This is also consistent with the earlier observation that in some cases, a normal and sclerotic involucrum may be indistinguishable among observers. Statistically, this is supported by the fact that the k value of Beit CURE (even with subtypes) is still consistently higher than that of the Penny classification.

This study did not aim to determine the effectiveness of the classification systems in terms of prognostic outcomes. Also, both classification systems are relatively recent and maybe even unfamiliar to most orthopedic surgeons, necessitating instructions, or providing review materials from the researchers. The lack of prior experience in using these classification systems may affect reliability outcomes. However, we can also use this to assess the ease with which these systems can be taught and retained.

CONCLUSION

In our study, the Beit CURE classification for pediatric CHOM had higher inter- and intraobserver reliability rates compared to the Penny classification. Further improvement in reliability can be made by combining B2 and B3 subtypes under the Beit CURE classification. Combining these two subtypes will result in better agreement between surgeons without compromising differences in treatment since both B2 and B3 subtypes will likely need a bone window or trough in addition to a sequestrectomy. The B1 subtype will still be addressed by localized removal of the sequestrum, while B4 would entail more reconstructive procedures to regain bone integrity after sequestrectomy or resection of non-viable bone. We agree with the original authors that a separate classification for physeal damage should be made.

We recommend that prospective studies be done using this modified Beit CURE classification for prognostic purposes. It would also be beneficial to our institution to provide a forum for educating surgeons and radiologists on this classification system to improve the inter- and intraobserver reliability further as this is still a relatively new classification.

Acknowledgment

The authors would like to express their gratitude to the Department of Orthopedics, including former chairman Dr. Edward Wang for his steadfast support in the name of orthopedic research, our Pediatric Orthopedic Consultants, and the residents for their invaluable contributions to this study.

The authors would also like to thank Dr. Esmeralda Cosette Atutubo for her support in making this study possible.

Lastly, the authors would like to thank the rest of the research staff for their invaluable assistance in this study.

Statement of Authorship

Both authors participated in paper writing, data collection and analysis, and approved the final version submitted. Dr. Javier provided mentorship and guidance throughout the process.

Author Disclosure

Both authors declared no conflicts of interest.

Funding Source

This paper was funded by the authors. No external funding from any public or commercial agency was declared by the authors.

REFERENCES

1. Jones HW, Harrison JW, Bates J, Evans GA, Lubega N. Radiologic classification of chronic hematogenous osteomyelitis in children. J Pediatr Orthop 2009;29(7):822-7. doi: 10.1097/BPO.0b013e3181b76933.
2. Delos Reyes CA, Ponio SS. An Epidemiologic Investigation of Chronic Osteomyelitis among Pediatric Patients Admitted from 2006 to 2010 at the Philippine General Hospital. Pediatr Infect Dis Soc Philipp J. 2013 Jan-Jun;14(1):14-23.
3. Lima AL, Oliveira PR, Carvalho VC, Cimerman S, Savio E. Recommendations for the treatment of osteomyelitis. Braz J Infect Dis. 2014;18(5):526–34. doi: 10.1016/j.bjid.2013.12.005.
4. Gosselin RA, Spiegel DA, Foltz M. Global Orthopedics: Caring for Musculoskeletal Injuries and Conditions in Resource-Poor Settings. New York: Springer-Verlag; 2014.
5. Romano CL, Romano D, Logoluso N, Drago L. Bone and joint infections in adults: a comprehensive classification proposal. Eur Orthop Traumatol 2011; 1(6):207–17. doi: 10.1007/s12570-011-0056-8.
6. Marais LC, Ferreira N, Aldous C, Roux TLB. The Classification of Chronic Osteomyelitis. SA Orthopaedic Journal Autumn 2014; 13 (1):22-28.
7. Spiegel DA, Penny JN. Chronic Osteomyelitis in Children. Techniques in Orthopaedics 2005; 20(2):142–52.
8. Beckles VL, Jones HW, Harrison WJ. Chronic haematogenous osteomyelitis in children: a retrospective review of 167 patients in Malawi. J Bone Joint Surg [Br] 2010;92-B:1138-43.
9. Jones HW, Beckles VLL, Akinola B, Stevenson AJ, Harrison WJ. Chronic haematogenous osteomyelitis in children: an unsolved problem J Bone Joint Surg [Br] 2011; 93-B: 8, 1005-1010.
10. Stevenson AJ, Jones HW, Chokotho LC, Beckles VL, Harrison WJ. The Beit CURE Classification of Childhood Chronic Haematogenous Osteomyelitis—a guide to treatment. J Orthop Surg Res 2015;10:144. doi: 10.1186/s13018-015-0282-9.
11. Patel AA, Vaccaro AR, Albert TJ, Hilibrand AS, Harrop JS, Anderson DG, et al. The Adoption of a New Classification System: Time-Dependent Variation in Interobserver Reliability of the Thoracolumbar Injury Severity Score Classification System. Spine 2007; 32(3): E105–E110. doi: 10.1097/01.brs.0000254107.57551.8a.
12. Viera AJ, Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. Fam Med 2005 May;37(5):360-3.