Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–7

DOI: 10.1177/20552076241277672
journals.sagepub.com/home/dhj

Sage

An assessment of the inter-rater and intra-rater reliability of the modified Gordon pin infection classification system

Anirejuoritse Bafor[1], Regitze Gyldenholm Skals[2], Ming Shen[3], Christopher A Iobst[1], Ole Rahbek[4], Søren Kold[4] and Marie Fridberg[4]

Abstract

Objectives: A grading system deployed for continuous at-home monitoring of pin sites would potentially increase the chances of early detection of pin-site infections and the commencement of early treatment. The first five grades of the Modified Gordon Pin Site Classification Scheme (MGS) meet the criteria for a visual-only, digital assessment-based grading system. The aim of this study was to assess the inter- and intra-rater reliability of the first five grades of the MGS from digital images.

Methods: We graded 1082 pin sites from 572 digital photographs of patients who underwent external fixator treatment for various conditions using the first five grades of the MGS classification scheme. Percent agreement and kappa values were calculated to determine the inter- and intra-rater agreement. Results were also grouped into two categories: "good" consisting of MGS grades 0–2 and "bad" made up of grades 3 and 4 for sensitivity analysis. We also analyzed reliability based on color only using MGS grades 0 and 2.

Results: A total of 843 of the 1082 pin sites were scored by all raters. There was moderate reliability between raters with a Fleiss kappa value of 0.48 [CI 0.45, 0.51]. The reliability remained moderate based on grouping into "good' versus "bad" and based on color with Fleiss kappa values of 0.48 [CI 0.45, 0.52] and 0.45 [CI 0.42, 0.49], respectively. Intra-rater reliability demonstrated substantial agreement with kappa values of 0.63.

Conclusion: Scoring pin sites from digital images with the MGS demonstrated only moderate inter-rater reliability. Modifying the use of digital photos is needed for at-home monitoring of pin sites.

Keywords

Modified Gordon score, inter-rater reliability, pin-site infection, pin-site labeling

Submission date: 10 January 2024; Acceptance date: 8 August 2024

Introduction

Pin site infections (PSIs) following the application of external fixators for fracture care or limb reconstruction procedures continue to be an important source of morbidity.[1,2] While desirable, a universally accepted consensus in the classification or grading of PSIs remains elusive, probably because there is also no consensus on the definition of a

[1]Center for Limb Lengthening and Reconstruction, Department of Orthopaedics, Nationwide Children's Hospital, Columbus, OH, USA
[2]Aalborg University Hospital, Research Data and Biostatistics, Aalborg, Denmark
[3]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
[4]Interdisciplinary Orthopaedics, Department of Orthopaedics, Aalborg University Hospital, Aalborg, Denmark

Corresponding author:
Anirejuoritse Bafor, Center for Limb Lengthening and Reconstruction, Department of Orthopedics, Nationwide Children's Hospital, Columbus, OH 43205, USA.
Email: anirejuoritse.bafor@nationwidechildrens.org

PSI.[3,4] This also explains the wide incidence of PSI reported in the literature.[5] This lack of agreement relates to the varied criteria used to diagnose PSIs, including clinical, radiological, and microbiological parameters.

Many authors have emphasized the need for a universal classification system for PSI to standardize reporting and for research purposes. This has driven research to refine the classification and grading of PSI. Several classification systems have been reported in the literature.[6–13] These classification schemes grade PSIs, for the most part, in order of increasing severity, using a combination of clinical symptoms and signs, radiological images, and microbiological assessments, including response to treatment. Besides being diagnostic, some of them also inform the choice of treatment.[7,9] The subjective nature of symptoms such as pain, swelling, and erythema, as well as the need for radiographic assessment, adds another layer of complexity to the application of many of these classification systems.

A classification scheme that relies solely on the visual appearance of the pin site would be an ideal resource for rapid screening to detect signs of infection. With the wide availability of smartphones that have a digital camera and the increasing use of telemedicine as a tool for patient review, the development of a classification system that is quick and easy to use without the need for radiological or microbiological assessment will be a welcome adaptation of the use of this technology. One of the advantages of this will be the potential for continuous, home-based self-monitoring and early diagnosis of PSI, as well as the prompt institution of treatment, which reduces the risk of more severe complications such as osteomyelitis.[14] It will also reduce the need for unnecessary in-person clinic visits for pin site assessment.

The Gordon score is a six-grade classification system for PSI that relies on the presence of pain, findings on plain radiographs, and the appearance of erythema, serous or purulent discharge.[10] Rahbek et al.[15] modified this classification system by eliminating the subjective symptom of pain and including a separate grade for the presence of serous drainage only. We aim to develop an image-only recognition algorithm that does not rely on subjective reporting of pain, plain radiographs, or microbiology results to facilitate home surveillance and early detection of PSI. We have thus further modified the classification system of Rahbek et al.[15] to rely only on the first five grades of the scoring system (the visible clinical signs). Since the appearance of radiolysis, sequestrum, and osteomyelitis are late events in the progression of PSI, eliminating these findings from the classification system will not invalidate our plan for a clinical image-only algorithm for the early detection of PSI. The rating of these visible clinical signs is based on subjective judgment and the reliability of judging digital images is unknown. Therefore, this study's goal was to evaluate the intra- and inter-rater reliability of the first five grades (the visible clinical signs) of the modified Gordon pin site classification system or "modified Gordon score (MGS)" in the grading of PSIs using a dataset of high-quality digital images of pin sites.

Methods

This was a retrospective multicenter study approved by the institutional review board (IRB) of Nationwide Children's Hospital, Columbus, Ohio. A waiver of consent for this study was sought and obtained from our IRB. We selected 1500 de-identified digital images of pin sites consecutively obtained using a fifth-generation Apple iPad (Version 16.7.5) with an 8-megapixel camera dedicated for this purpose as part of the routine standard of care from patients who had undergone application of an external fixator for various reasons between the 1st of January and the 31st of December 2021 for inclusion in this study.

The images were evaluated by a single author and divided into two groups based on quality: high and low. Only images of pin sites assessed as "high quality," showing clearly visible skin around the pin or wire, with the view not obstructed by the fixator, of acceptable quality, and not blurred, were included in this study. Five hundred and seventy-two (572) images were deemed high quality. We sequentially numbered the images for easy identification. Each pin site was annotated and numbered systematically (Figure 1). The identification number for each pin site included the image number, the number of pin sites in the image, and the specific pin site number. Thus, for example, a pin site would be identified as "IMG_0090_3_2" where "IMG_0090" was the image number, "3" was the total number of pin sites located in the image, and "2" referred to the pin site labeled as "2" in image number "IMG_0090." A total of 1082 pin sites from 572 images which included different types of skin color across the spectrum of skin colors were finally available for this study.

The MGS classification system (Table 1) was used to grade pin sites.[15] Because this study did not rely on findings on plain radiography for grading, we scored the pin sites from 0 to 4 as per the first five grades of the modified Gordon pin infection classification (the visual grades).

Two orthopedic surgeons (MF and AB) and an experienced orthopedic nurse (TJ) with a mean of 12.3 years of clinical experience and blinded to each other's evaluation graded all 1082 pin sites using the MGS. The raters had a pre-assessment meeting to review and agree on modalities for grading to create some standardization of the assessment method. Each rater received a Microsoft Excel spreadsheet containing details of all images and annotated pin sites to record their classification of the pin sites and access to a Dropbox folder containing all the selected images. For pin sites that could not be graded for some reason, raters were instructed to enter "not applicable" (NA) into the spreadsheet. All other pin sites received a MGS based on the classification system.

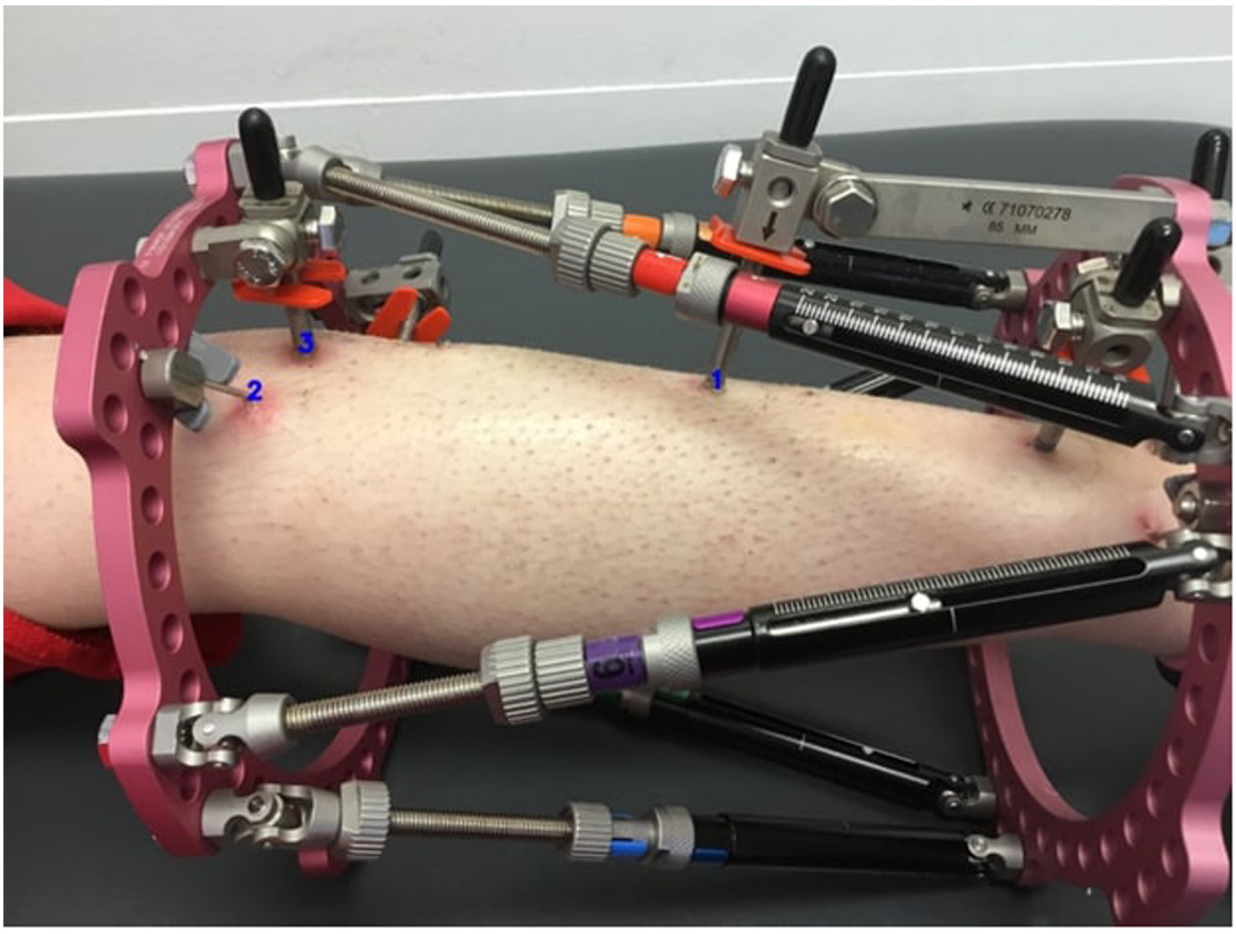

Figure 1. Sample image showing annotated pin sites.

Table 1. The modified Gordon pin site infection classification.[15]

Grade	Description
0	Clean
1	Serous drainage, no erythema
2	Erythema, no drainage
3	Erythema and serous drainage
4	Erythema and purulent drainage
5	Erythema, purulent drainage, radiographic osteolysis
6	Ring sequestrum or osteomyelitis

For sensitivity analysis, we divided the visual MGS into two groups. The first three grades (0, 1, and 2) were classified as "good," and the last two grades (3 and 4) were classified as "bad." For a "color analysis," we also determined the inter-rater reliability of all pin sites graded as either 0 or 2, with 0 representing normal-looking skin around the pin site and 2 representing erythematous skin.

Four months later, a statistician, independent of the raters, randomly selected 100 pin-site images from the pool of 1082 pin sites. The two orthopedic surgeons independently assessed these images using the modified Gordon pin infection classification to evaluate intra-rater reliability.

Statistical analysis

Data were analyzed using *R* statistical software version 4.2.2. Percent agreement as well as kappa values were used to estimate intra- and inter-rater reliability. Confidence intervals for the agreement were calculated using the Clopper–Pearson method for binomial proportions. Fleiss kappa was chosen to calculate inter-rater reliability between the three raters because the Fleiss kappa statistic considers the possibility of chance in the reliability of this classification system. Cohen's kappa values with

95% confidence intervals for categorical data were calculated to estimate the pairwise intra-rater reliability. The prevalence and bias-adjusted kappa (PABAK) values were also calculated for sensitivity analysis of the data set. PABAK values are useful in the reliability assessment of skewed distribution of data where the prevalence rates in one or more groups are high.[16,17] The interpretation of results was based on the recommendations of Landis and Koch.[18] Thus, for kappa values of <0, the interpretation was that there was no agreement; for 0 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement, and 0.81 to 1.0 was considered perfect agreement.

Results

Of the 1082 pin sites, 843 were scored by all three raters, and only these were included in the analysis for reliability. A total of 239 pin sites were not assessed by all three raters for reasons ranging from poor light reflection to uncertainty about the nature of fluids noticed around the pin sites (see Figure 2 for a flowchart).

There was moderate inter-rater reliability between the three raters with a Fleiss kappa value of 0.48 [CI 0.45, 0.51]. The results were similar when we grouped pin sites into "good" (MGS scores 0–2) and "bad" (MGS scores 3 and 4) categories with a Fleiss kappa value of 0.48 [CI 0.45, 0.52]. For color analysis, where only pins graded as 0 or 2 were isolated and rated, we found slightly lower but still, moderate inter-rater reliability with a Fleiss kappa value of 0.45 [CI 0.42, 0.49]. Pairwise inter-rater reliabilities are reported in Table 2. Compared to these findings, sensitivity analyses for the "good" versus "bad" and the "0 versus 2" color analysis groups, respectively, showed PABAK values ranging from 0.80 to 0.92 and from 0.41 to 0.57, respectively. For the MGS scores, the PABAK values ranged from 0.38 to 0.56. The percentage agreement was 63.1% among all three raters and rose to 88.7% when pin sites were grouped as "good" versus "bad." Results are shown in Table 2.

On the assessment of the individual classifications, the inter-rater reliability was best for pin sites classified as MGS 0, showing moderate reliability with a Fleiss kappa of 0.57. When pin sites were grouped as either "good"

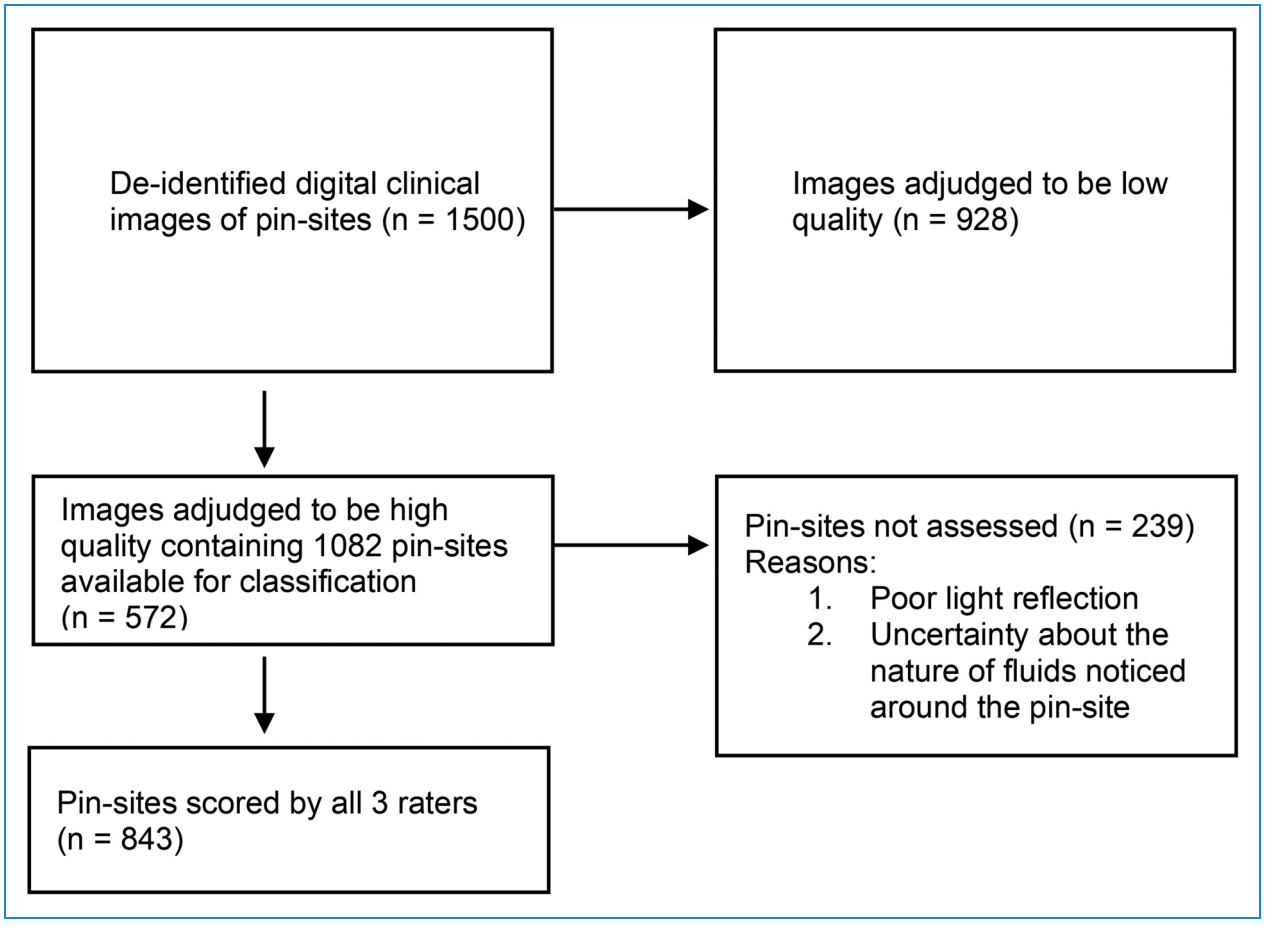

Figure 2. Schematic diagram showing selection criteria for images graded using the modified Gordon pin-site classification scheme.

Table 2. Inter-rater reliability of pin-site infection assessment using the modified Gordon pin-site classification system (MGS).

	MGS score 0–4		Grouped MGS score (0–2 vs. 3, 4)		Color grouping (MGS grade 0 vs 2)	
Inter-rater	Kappa	% agreement	kappa	% agreement	Kappa	% agreement
All raters	0.48 [0.45, 0.51]	63.1%	0.48 [0.45, 0.52]	88.7%	0.45 [0.42, 0.49]	64%
1 versus 2	0.52 [0.46, 0.57]	77.1%	0.39 [0.29, 0.49]	89.8%	0.49 [0.43, 0.55]	77.8%
3 versus 2	0.40 [0.35, 0.45]	68.9%	0.54 [0.44, 0.63]	91.6%	0.36 [0.30, 0.42]	70.4%
3 versus 1	0.53 [0.48, 0.59]	78.2%	0.58 [0.45, 0.71]	96.1%	0.50 [0.44, 0.56]	78.3%

MGS: modified Gordon score.

(MGS 0, 1, or 2) or "bad" (MGS 3 or 4), the inter-rater reliability was moderate, with a Fleiss kappa of 0.48 for both groups. The individual reliability assessment when pin sites were grouped based on color alone (MGS 0 or 2) was moderate for both groups, with a Fleiss kappa of 0.56 and 0.44 for MGS 0 and 2, respectively.

Intra-rater reliability testing on 100 randomly and independently selected pin site images demonstrated substantial intra-rater agreement, with both raters demonstrating kappa values of 0.63 [CI 0.47, 0.80] for rater 1 and 0.63 [CI 0.48, 0.79] for rater 2.

Discussion

The intermittent nature of in-person assessments following the application of external fixators creates delays in the diagnosis and treatment of pin-site infections which can potentially have serious consequences. The prospect of continuous at-home monitoring, especially with the increasing use of virtual consultations and telemedicine, creates a promising opportunity for the early detection of PSIs. This is even more relevant with the growing trend toward the use of machine learning and artificial intelligence (AI) algorithms in modern-day medicine to aid the diagnosis of disease conditions and the early detection of complications. Continuous at-home monitoring is best achieved using a screening tool that offers rapid assessment of the visual appearance of the pin site since some of the early clinical signs of infection, such as erythema, are diagnosed visually. Utilizing the first five grades of the MGS meets the criteria for a visual-only scheme to screen pin sites for infection. Indeed, the original classification by Gordon et al.[10] did not report findings of osteolysis (grade 4) or osteomyelitis (grade 5) in 4473 pin site evaluations in their series, further justifying our choice of excluding these variables in our modified classification scheme. The MGS has not been previously validated by reliability studies, which makes the findings of our study important. We assessed the reliability of the visual-only clinical component of the modified Gordon pin-site classification system, including several modifications to simplify the classification system further. Our results demonstrated moderate inter-rater and substantial intra-rater reliability of the classification system.

Several pin site grading systems have been tested for inter- and intra-rater reliability with varying results. The Checkett's grading system demonstrated good intra-rater and poor to moderate inter-rater reliability in the grading of pin sites.[19] In that study, participants highlighted the difficulty in assessing a pin site based solely on a photograph. They highlighted the importance of clinically elicited symptoms and signs to improve diagnostic accuracy and confidence in grading. In contrast, Clint et al.[11] reported excellent inter- and intra-rater agreement between two raters when using a clinical grading system that included pain as a variable in a relatively small cohort of 15 patients, with a total of 218 pin sites. These contrasting findings highlight the differences in assessing pin sites in person versus assessment from photographs. Fridberg et al.[20] reported a 98% observed agreement using the MGS in a clinical setting, compared to our finding of 63.1% agreement when the assessment was carried out on digital photographs. Even with the exclusion of pain as a variable in the classification system, the MGS performed better when applied clinically compared to an assessment from digital photographs. The likely reason is the possibility of a three-dimensional visualization of a pin site during an in-person visit. This provides a better chance of assessing the pin site compared to the two-dimensional nature of digital photographs. The pin site can be viewed from multiple angles, and the effect of ambient lighting is also considered; thus, any doubts regarding image visualization or perception are easily dispelled. The fact that 239 pin sites in this study were not assessed by all three raters further supports this notion.

The inter-rater agreement is also affected by the subjective perception of the signs of inflammation around a pin site and the determination of the cause (response to the pin or infection). Iliadis et al.[3] highlighted this subjectivity in a

systematic review of current classification schemes for PSIs. Pin site pain was not an included variable in our classification scheme as we sought to assess a scale that relied solely on the visual appearance of the pin site. Pain is, however, one of the cardinal symptoms of inflammation, and its inclusion may improve the diagnostic accuracy of any reliable classification scheme. It is also possible that the absence of pain as a variable in the MGS score contributed to the finding of only substantial intra-rater and moderate inter-rater reliability.

Out of 1500 digital images initially selected in the current study, only 572 images were found to have sufficient quality for pin-site classification. On these 572 images, a total of 1082 pin sites were available for classification, but 239 pin sites were not assessed by all raters due to reasons ranging from poor light reflection to uncertainty about the nature of fluids (serous or purulent) noticed around the pin sites. This highlights one of the limitations of this study inherent in its retrospective design**.** Even in the remaining optimized pin-site images, which were classified by the raters, only moderate inter-rater and intra-rater agreements were found, highlighting the limitation of a visual-only system based on digital images for assessing PSIs. One option to address this problem might be to incorporate the use of multiple images of the same pin site taken in at least two different planes as is done with plain X-rays. This provides multiple views of the same pin site, which in turn offers enhanced visual feedback and thus addresses situations where doubt exists. In addition, the lack of verbal patient feedback relating to variables such as pain diminishes the possibility of diagnosing pin-site infections. However, since one of our long-term goals is to develop a screening tool to determine which patients and pin sites require further evaluation, including pain either as an independent variable or as in the original classification by Gordon et al.,[10] might improve the reliability and utility of this classification system as a screening tool for PSI. The subjectivity of pain is one of the limitations of its reliability, but this may be mitigated by qualifying characteristics such as the recent onset of pain or the presence of increasing intensity of pain at a localized site. There have also been promising results with thermography as an adjunct to assessing PSI.[15,21] Including thermography might further increase the diagnostic accuracy and reliability of pin site classification schemes.

Another limitation of this study is that most pin sites assessed were either MGS grade 0 or 2. The kappa statistic is based on marginal distributions of categorical or ordinal data, which account for the measure of chance agreement between raters. The skewed nature of the dataset creates an imbalance, which unveils one of the weaknesses of the kappa statistic.[17] This accounts for the wide variation between the kappa value and the percentage agreement when comparing the same level of reliability testing. It can also cause unexpectedly low kappa values. This weakness was most pronounced for the grouped MGS score. The PABAK values which are designed to address this limitation did not vary much from the kappa values for the MGS score in this study, however, we did notice a remarkable change in the PABAK values when scores were grouped into "good" versus "bad." This is important to the extent that the PABAK values indicate the effect of prevalence and bias on the true kappa values and should thus be considered when interpreting the results of reliability testing.[16] A further limitation is the difficulty in differentiating the initial signs of a PSI from physiologic reactions to a pin or wire, which may arise from a foreign body skin reaction represented by local irritation and normal physiologic scarring. This problem is also related to the lack of a consensus on defining a PSI. This creates problems in deciding on the most appropriate form of treatment required. More importantly, it may also lead to the inappropriate use of antibiotics in these patients. While it might remain inevitable that early PSI may be indistinguishable from local physiological irritation, continuous monitoring provides the opportunity for early detection and the institution of prompt and appropriate treatment.

Conclusion

Our study revealed an inter-rater reliability of 0.48 (Fleiss kappa) and an intra-rater reliability of 0.63 using the MGS to grade pin sites. Grouping pin sites into simplified grades or based on the presence or absence of erythema did not change the reliability of the classification scheme much. These findings represent suboptimal results of reliability testing of a grading system intended to be the basis for an AI algorithm. Digital images alone are thus not enough to diagnose and grade PSI.

Acknowledgements: The authors thank nurse Trine Lyngholm Jensen for her help in scoring pin sites.

Contributorship: AB contributed to conceptualization, data curation, formal analysis, investigation, methodology, writing (original draft), and writing (review and editing). RGS contributed to data curation, formal analysis, investigation, methodology, and writing (review and editing). MS contributed to data curation, software, formal analysis, methodology, resources, and writing (review and editing). CI, OR, and SK contributed to conceptualization, supervision, and writing (review and editing). MF contributed to conceptualization, data curation, software, formal analysis, investigation, methodology, resources, writing (original draft), and writing (review and editing).

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The IRB at the Research Institute, Nationwide Children's Hospital, Columbus, Ohio, USA approved this study (STUDY00001871).

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Guarantor: Anirejuoritse Bafor.

ORCID iDs: Anirejuoritse Bafor https://orcid.org/0000-0001-9278-5324
Ole Rahbek https://orcid.org/0000-0002-5602-4533
Søren Kold https://orcid.org/0000-0002-3387-1473

References

1. Iobst C and Liu R. A systematic review of incidence of pin track infections associated with external fixation. *J Limb Lengthening Reconstr* 2016; 2: 6–16.
2. Jauregui JJ, Bor N, Thakral R, et al. Life-and limb-threatening infections following the use of an external fixator. *Bone Jt J* 2015; 97-B: 1296–1300.
3. Iliadis AD, Shields DW, Jamal B, et al. Current classifications of pin site infection and quality of reporting: A systematic review. *J Limb Lengthening Reconstr* 2022; 8: S59–S68.
4. Santy J. A review of pin site wound infection assessment criteria. *Int J Orthop Trauma Nurs* 2010; 14: 125–131.
5. Iobst CA. Pin-track infections: Past, present, and future. *J Limb Lengthening Reconstr* 2017; 3: 78–84.
6. Paley D. Problems, obstacles, and complications of limb lengthening by the Ilizarov technique. *Clin Orthop Relat Res* 1990; 250: 81–104.
7. Checketts R, Otterburn M and MacEachern G. Pin track infection: Definition, incidence and prevention. *Int J Orthop Trauma* 1993; 3: 16–18.
8. Dahl MT, Gulli B and Berg T. Complications of limb lengthening: A learning curve. *Clin Orthop Relat Res* 1994; 301: 10–18.
9. Checketts R, MacEachem A and Otterburn M. Pin track infection and the principles of pin site care. In: De Bastiani G, Apley A and Goldberg A (eds) *Orthofix external fixation in trauma and orthopedics*. London: Springer, 2000, pp.97–103.
10. Gordon JE, Kelly-Hahn J, Carpenter CJ, et al. Pin site care during external fixation in children: Results of a nihilistic approach. *J Pediatr Orthop* 2000; 20: 163–165.
11. Clint SA, Eastwood DM, Chasseaud M, et al. The 'good, bad and ugly' pin site grading system. A reliable and memorable method for documenting and monitoring ring fixator pin sites. *Injury* 2010; 41: 147–150.
12. Chan CK, Saw A, Kwan MK, et al. Diluted povidone-iodine versus saline for dressing metal-skin interfaces in external fixation. *J Orthop Surg (Hong Kong)* 2009; 17: 19–22.
13. Santy-Tomlinson J, Vincent M, Glossop N, et al. Calm, irritated or infected? The experience of the inflammatory states and symptoms of pin site infection and irritation during external fixation: a grounded theory study. *J Clin Nurs* 2011; 20: 3163–3173.
14. Ceroni D, Grumetz C, Desvachez O, et al. From prevention of pin-tract infection to treatment of osteomyelitis during paediatric external fixation. *J Child Orthop* 2016; 10: 605–612.
15. Rahbek O, Husum HC, Fridberg M, et al. Intrarater reliability of digital thermography in detecting pin site infection: A proof of concept study. *Strateg Trauma Limb Reconstr* 2021; 16: 1–7.
16. Sim J and Wright CC. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys Ther* 2005; 85: 257–268.
17. Flight L and Julious SA. The disagreeable behaviour of the kappa statistic. *Pharm Stat* 2015; 14: 74–78.
18. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
19. Groenewoud R, Chhina H, Bone J, et al. Inter-and intra-rater reliability of the Checketts' grading system for pin-site infections across all skin colours. *Strateg Trauma Limb Reconstr* 2023; 18: 2–6.
20. Fridberg M, Ghaffari A, Husum HC, et al. Evaluating interrater reliability of the modified Gordon score for pin site infection. *Orthop Proc* 2023; 105-B: 8.
21. Annadatha S, Hua Q, Fridberg M, et al. Preparing infection detection technology for hospital at home after lower limb external fixation. *Digit Heal* 2022; 8: 20552076221109504.