DOI: https://doi.org/10.30795/jfootankle.2025.v19.1938

Original Article

Can posterior malleolus fracture classification influence treatment choice?

Bruno Abdo Santana de Araújo[1], Saulo Pereira de Oliveira[1], João B. Ferreira Junior[2], Rafael Santini da Silva[2], Márcio Oliveira[3], Henrique Mansur[1]

1. Hospital Santa Helena, Brasilia, DF, Brazil.
2. Instituto Federal do Sudeste Minas Gerais - Campus Rio Pomba, Lindo Vale, Rio Pomba, MG, Brazil.
3. Centro de Reabilitação Return to Play, Brasília, DF, Brazil.

Abstract

Objectives: To analyze whether PMF classifications can influence treatment choice and surgical approach. In addition, verify the intra- and interobserver reproducibility of the three main classifications, stratified by observers' experience.

Methods: Ankle computed tomography of 50 patients was evaluated by ten observers, four orthopedists specialized in foot and ankle surgery, and six non-specialist orthopedists, with an interval of two weeks between evaluations. The evaluators classified PMF according to the Mason, Haraguchi, and Bartoníček/Rammelt classifications and determined whether to treat PMF conservatively or surgically (in this case, by access route). In addition, the reproducibility of the classifications was evaluated. The association between decision-making and access route was analyzed using the Chi-Square Test (χ^2). Cronbach's alpha was used to assess intraobserver agreement, and the kappa statistic was used to evaluate interobserver agreement.

Results: In analyses of decision-making and the access route, all classifications showed large effect sizes (V > 0.50). Intraobserver reproducibility across the entire sample ranged from 0.53 to 0.95 (0.78 ± 0.12) for the Haraguchi classification, from 0.47 to 0.95 (0.74 ± 0.17) for Mason, and from 0.53 to 0.94 (0.72 ± 0.12) for Bartoníček/Rammelt, indicating adequate agreement across the three classifications. For the specialist orthopedists, the mean ratings for the Haraguchi, Mason, and Bartoníček/Rammelt classifications were 0.86, 0.84, and 0.75, corresponding to good, good, and adequate, respectively. For the group of non-specialists, the means were 0.72, 0.68, and 0.70, indicating adequate, average, and adequate, respectively. Interobserver reproducibility was considered reasonable for Haraguchi (0.38) and moderate for Mason (0.42) and Bartoníček/Rammelt (0.43).

Conclusion: All three classifications had large effects on treatment choice and access route decisions, with the Bartoníček/Rammelt classification showing the highest effect. All three PMF classifications were considered adequately reproducible by intraobserver assessment. Interobserver reproducibility was considered reasonable for Haraguchi and moderate for the others.

Level of evidence IV; Therapeutic studies; Case series.

Keywords: Classification; Ankle fractures; Reproducibility of results.

Introduction

Ankle fractures account for 9% of all fractures in adults, with an estimated incidence of 1 fracture per 1,000 inhabitants per year[1]. Among these, posterior malleolus fractures (PMF) account for up to 44% and are associated with a worse prognosis[2].

The timing of PMF repair remains a subject of debate in the literature and a source of ongoing uncertainty among surgeons. A recent study found no difference in long-term functional and radiological outcomes[3], although patients who underwent PMF fixation had larger fragments than those treated conservatively. In addition to the size of the

Study performed at the Hospital Santa Helena, Brasilia, DF, Brazil.

Correspondence: Bruno Abdo Santana de Araújo. SHLN 516 Conjunto D, Asa Norte, 73015132, Brasilia, DF, Brazil. **Email:** brunoabdosa@gmail.com. **Conflicts of interest:** none. **Source of funding:** none. **Date received:** September 10, 2025. **Date accepted:** December 11, 2025.

How to cite this article: Araújo BAS, Oliveira SP, Ferreira Junior JB, Silva RS, Oliveira M, Mansur H. Can posterior malleolus fracture classification influence treatment choice? J Foot Ankle. 2025;19(3):e1938.

posterior malleolus fragment, a factor previously considered the most important, fracture morphology has been used as a decisive parameter in decision-making, particularly on computed tomography (CT)[4]. The most commonly used PMF classifications, based on CT, are Mason et al.[5], Haraguchi et al.[6], and Bartoníček et al.[7]

Fracture classification systems aim to characterize the injury, guide treatment, and indicate prognosis, in addition to facilitating communication among surgeons and organizing knowledge for its incorporation into clinical and epidemiological databases. A good classification must be validated and reliable, and exhibit high inter- and intraobserver reproducibility[8,9].

Previous studies evaluated the inter- and intraobserver reproducibility of these three PMF classifications, yielding similar results; however, they did not determine which was most reproducible and employed different methodologies[10-12]. Interestingly, none of these studies examined whether these classifications influence the choice of PMF treatment—surgical vs conservative—an important criterion for a good classification. In addition, these studies included observers with varying levels of specialization, from medical students and orthopedic residents to specialists in ankle and foot surgery and traumatologists, information that may have directly influenced the results due to the lack of experience among some participants.

The objective of this study is to analyze whether PMF classifications can influence treatment choice and surgical approach. In addition, verify the intra- and interobserver reproducibility of the three main classifications, stratified by observers' experience.

Methods

This is a cross-sectional repeated-measures study, following the recommendations of GRASS[13] and approved by the Institutional Review Board. All cases of ankle fractures involving the posterior malleolus at a single institution from October 2021 to December 2023 were included. Cases of fractures in patients with an immature skeleton or with an incomplete radiological study were excluded.

The diagnosis of PMF was based on imaging, including radiographs and CT of the injured ankle. Using CT images in axial and sagittal views, an online questionnaire was created in video format (https://forms.gle/PBt4Lgy1gs2a1DS87), containing 50 cases of PMF, for the evaluation of ten independent observers, all orthopedists qualified by the Brazilian Society of Orthopedics and Traumatology, four subspecialists in ankle and foot surgery, and six non-specialist orthopedists (without other subspecialties). All participants were trained in each PMF classification before completing the questionnaire. Participants classified each case according to the three PMF classifications: Haraguchi et al.[6] (types 1, 2 or 3); Mason et al.[5] (types 1, 2A, 2B or 3); and Bartoníček et al.[7] (types 1, 2, 3, 4, or 5), on two occasions, with an interval of two weeks between them. They also addressed treatment options for the posterior fragment, regardless of the presence of other injuries, and, if the posterior malleolus was to be fixed, which approach they would use ("percutaneous", "posteromedial open", or "posterolateral open").

Statistical analysis

The descriptive analysis presented the observed data in tables. Cronbach's alpha was calculated to assess internal consistency among the items evaluated. The sample and measured items were selected to reflect a single evaluation task performed repeatedly by a single evaluator (intraobserver reproducibility). Interpretation intervals for Cronbach's alpha used were: $\alpha \geq 0.9$: Very good to excellent internal consistency (with stronger interpretation of reproducibility, suitable for accurate measurements); $\alpha \geq 0.9$: Excellent (high reproducibility); $0.8 \leq \alpha \leq 0.9$: Good (good reproducibility); $0.7 \leq \alpha \leq 0.8$: *Adequate (acceptable reproducibility); 0.6* $\leq \alpha \leq$ *0.7: Average (questionable reproducibility);* $\alpha \leq$ *0.6: Low (unsatisfactory reproducibility)*[14].

The Kappa test (κ) was used to assess interobserver agreement. This test measures the degree of agreement among evaluators beyond what is expected by chance. To classify the results of the Kappa test, the following parameters were used: Kappa test (κ) interobserver agreement, $\kappa \leq 0.20$: Poor agreement; $0.21 \leq \kappa \leq 0.40$: Reasonable agreement; $0.41 \leq \kappa \leq 0.60$: Moderate agreement; $0.61 \leq \kappa \leq 0.80$: Substantial agreement; and $\kappa \geq 0.81$: Almost perfect agreement[15].

The Chi-square test (Q^2) was applied to the tables "Decision / Method correlation" and "Access path / Method". When the Chi-square test assumptions (i.e., expected frequencies of at least 5%) were not met, Fisher's exact test was used. In addition, to enhance robustness, a Monte Carlo test was performed, yielding a more accurate estimate of statistical significance. Effect sizes were calculated and suitable for each matrix (2x3 "Decision / Method" and 3x3 "Access path / Method"). For table 3×2 (k=1), the effect size was considered: Small: V > 0.10, Medium: V > 0.30, Large: V > 0.50. For table 3×3 (k=1), the effect size was considered: Small: V > 0.07, Medium: V > 0.21, Large: V > 0.35[16].

The significance criterion adopted was the 5% level. Statistical analysis was performed using SAS System, version 6.11 (SAS Institute, Inc., Cary, North Carolina, USA).

Results

When analyzing decision-making based on classifications, with reference to the value of Phi Cramér's (V), all classifications showed a large effect size (V > 0.50), but with a higher absolute value in the Bartoníček/Rammelt classification (V = 0.72) than in Mason (V = 0.70) and Haraguchi (V = 0.69). Regarding the choice of access route, the three classifications showed large effect sizes (V > 0.35), with the Bartoníček/Rammelt classification having the largest absolute effect size (V = 0.40) compared to Maison and Haraguchi (V = 0.37 and 0.37, respectively).

The results of the intraobserver agreement, measured by Cronbach's alpha, and their interpretation are presented in detail in Table 1. The intraobserver agreement across the entire sample was 0.78 (0.12) for the Haraguchi classification, 0.74 (± 0.17) for Mason, and 0.72 (± 0.12) for Bartoníček/Rammelt, indicating adequate agreement. The mean intraobserver agreement in the expert group was 0.86 (± 0.09) for the Haraguchi classification, 0.84 (± 0.15) for Mason, and 0.75 (± 0.17) for Bartoníček/Rammelt. These results were considered reproducible, with good, good, and adequate reproducibility, respectively. Among non-specialists, the results were lower, with means of 0.72 (± 0.12), 0.68 (± 0.17), and 0.70 (± 0.08) for Haraguchi, Mason, and Bartoníček/Rammelt, which were considered adequate, average, and adequate, respectively.

Regarding the degree of reproducibility of classifications among observers (measured by the Kappa test), it was considered reasonable for Haraguchi ($\kappa = 0.38$) and moderate for Mason ($\kappa = 0.42$) and Bartoníček/Rammelt ($\kappa = 0.43$), all presenting statistical significance ($p < 0.05$).

Discussion

This study analyzed the influence of the three main PMF classifications on treatment decision-making and inter- and intraobserver reproducibility. The main findings were that the treatment choice (fix or do not fix the PMF) and the access route had substantial effects across the three classifications, with the absolute values being higher for the Bartoníček/Rammelt classification. Intraobserver agreement was considered adequate for the Haraguchi, Mason, and Bartoníček/Rammelt classifications. Separating by group, the agreement of the three classifications was considered good, good, and adequate among specialists in foot and ankle surgery, and adequate, average, and adequate among non-specialists. Interobserver reproducibility was considered reasonable for Haraguchi and moderate for Mason and Bartoníček/Rammelt.

A good classification, in addition to being reproducible, should also aid in the treatment[8,9]. To our knowledge, our study was the first to evaluate the agreement between PMF classifications and treatment decisions. We asked the observers to determine whether the analyzed fracture would be managed surgically or conservatively, and, if surgically, which access route to use. To define the conduct, regardless of the degree of specialization, the three classifications presented a large effect size, that is, all helped the observers in their therapeutic choice, with the Bartoníček/Rammelt classification having the highest absolute value, the one that most helped in the decision, and the Haraguchi classification having the lowest value. We believe that this result reflects the ease with which the Bartoníček classification[7] proposes the PMF treatment compared to the other two classifications: type 1 fractures are considered non-surgical treatment injuries, while all other four types are mostly surgical treatment. For Mason[5], the decision to operate or not depends more on syndesmosis stability tests than the fracture classification itself. Haraguchi[6], despite not having its criteria properly validated, suggests that type 1 fractures should only be addressed if they remain poorly reduced after fixation of the lateral and medial malleoli; type 2 fractures with two fragments, initially only the medial fragment is fixed, making the fracture a "type 1", then following the same criteria; if it is a type 2 fracture with only one fragment, its treatment must be surgical. Based on the observed results, we believe that the Bartoníček/Rammelt classification is most useful for defining the treatment, thereby facilitating the surgeon's decision.

In addition to the choice of PMF fixation, we investigated whether these classifications also help observers select the surgical access route. Thus, in the surgical cases, participants were asked to choose between "percutaneous access routes from anterior to posterior", "posteromedial access", or "posterolateral access". The three classifications showed large effect sizes, with the Bartoníček/Rammelt classification having the largest absolute effect size. We believe that this result occurred due to the ease with which the Bartoníček/Rammelt classification identifies the fracture trace and the main fragment. To our knowledge, no other study has analyzed the correlation between classifications and the choice of access route for the surgical approach. However, it is noteworthy that none of these classifications can accurately describe the complexity of PMF, since none of them considers the presence of a fragment interposed in the fracture focus, the degree of joint impaction, or the degree of deviation of the fragment[2], and the surgeon must perform a detailed study and surgical planning, based on imaging tests, especially CT.

Several factors can influence the reproducibility of a classification[9]. In our study, higher observer experience was

Table 1. Intraobserver agreement for each of the posterior malleolar fracture classifications, including maximum and minimum values (mean ± standard deviation).

	Haraguchi	Mason	Bartoníček/Rammelt
Overall (n = 10)	0.53 - 0.95 (0.78 ± 0.12) (α = adequate)	0.47 - 0.95 (0.74 ± 0.17) (α = adequate)	0.53 - 0.94 (0.72 ± 0.12) (α = adequate)
Specialists (n = 4)	0.75 - 0.95 (0.86 ± 0.09) (α = good)	0.63 - 0.95 (0.84 ± 0.15) (α = good)	0.53 - 0.94 (0.75 ± 0.17) (α = adequate)
Non-specialists (n = 6)	0.53 - 0.84 (0.72 ± 0.12) (α = adequate)	0.47 - 0.87 (0.68 ± 0.17) (α = average)	0.68 - 0.79 (0.70 ± 0.08) (α = adequate)

associated with greater intraobserver reproducibility in PMF classifications. According to the evaluations of non-specialist orthopedists, the reproducibility of the Haraguchi, Mason, and Bartoníček/Rammelt classifications was adequate, average, and adequate, respectively. Among the specialists, the classifications of Haraguchi and Mason showed greater reproducibility and were considered good, as indicated by Cronbach's alpha. The Bartoníček/Rammelt classification yielded similar values, indicating adequate reproducibility across levels of specialization. These results differ from previous studies, in which the observer's experience did not affect the reproducibility of these classifications[10-12]. This divergent finding can be explained by the substantial heterogeneity among the observers, who ranged from subspecialists to medical students, whereas our evaluators were all trained orthopedists, including four subspecialists. A recent study[17] found that among its observers, specialists in foot and ankle surgery achieved the highest intra- and interobserver reproducibility compared with non-subspecialist orthopedists in this area, orthopedic residents, and radiologists. Likewise, we believe that the greater the observer's experience, the easier it is to identify fracture traits, thereby making this analysis more consistent and supporting our results.

Interobserver reproducibility indicates the consistency of evaluations across individuals using the same instrument and is an important metric for assessing the validity of a classification[15,18]. The interobserver reproducibility in our study was lower than that reported in previous studies[10-12]. For the Haraguchi classification, our observers showed reasonable reproducibility (k = 0.38), whereas other studies reported moderate (10,11) or substantial (12) values. Regarding Mason's classification, reproducibility was comparable to that reported in other studies. We obtained moderate agreement (k = 0.42), which is close to that reported in previous studies[10-12]. The Bartoníček/Rammelt classification had the highest interobserver reproducibility in our study, with moderate reproducibility (k = 0.43), similar to that reported by Morales et al.[10] (k = 0.53). Other studies have also reported that this last classification has the highest interobserver reproducibility, with substantial agreement[11,12]. Collectively, we obtained divergent results from previous studies, which can be explained by differences in the evaluation methodologies employed. Our study included 50 cases of PMF and ten evaluators, who completed the questionnaire on two occasions, with an interval of two weeks between each. The other studies used 94 cases and six evaluators[10], 60 cases and nine evaluators[11], and 113 cases and four observers[12], with intervals between evaluations of three, four, and eight weeks, respectively. Previous reliability studies suggest approximately ten cases per observer to ensure adequate statistical power for inter- and intraobserver agreement analyses[13,15]. The observed results demonstrate that changes in the number of cases to be evaluated, the number of evaluators or evaluations, and the interval between them can influence the research findings.

Our study has several limitations, including the small sample size compared to previous studies. The choice of observers (orthopedists specializing in foot and ankle and non-specialists) introduces selection bias; however, this approach was used to address results already reported in the literature. Furthermore, unqualified observers (medical students or radiologists) are not expected to decide on the therapeutic modality, the main objective of this study. In addition, we did not provide the evaluators with three-dimensional CT reconstruction images, which may have affected the reproducibility of the evaluated classifications. A previous study showed that three-dimensional analysis of pylon fractures improves the understanding and reproducibility of their classification and preoperative planning[19], a factor that may be similar in PMF. It is important to highlight that the cross-sectional design of our study does not allow for establishing a causal relationship between classifications and therapeutic decisions or the choice of access route.

Conclusion

All three classifications had large effects on treatment choice and access route decisions, with the Bartoníček/Rammelt classification showing the highest effect. All three PMF classifications were considered adequately reproducible by intraobserver assessment. Interobserver reproducibility was considered reasonable for Haraguchi and moderate for the others.

Authors' contributions: Each author contributed individually and significantly to the development of this article: BASA *(https://orcid.org/0000-0001-5269-9106) Data collection, interpreted the results of the study, bibliographic review, wrote the article; SPO *(https://orcid.org/0000-0002-5649-8122) Data collection, interpreted the results of the study, bibliographic review, wrote the article; JBFJ *(https://orcid.org/0000-0002-7541-8212) and RSS *(https://orcid.org/0009-0001-3860-9116) Statistical analysis; MO *(https://orcid.org/0000-0003-3147-5925) ethical approval; HM *(https://orcid.org/0000-0001-7527-969X) Data collection, interpreted the results of the study, bibliographic review, wrote the article. All authors read and approved the final manuscript. *ORCID (Open Researcher and Contributor ID).

References

1. Court-Brown CM, Caesar B. Epidemiology of adult fractures: A review. Injury. 2006;37(8):691-7.
2. Terstegen J, Weel H, Frosch KH, Rolvien T, Schlickewei C, Mueller E. Classifications of posterior malleolar fractures: a systematic literature review. Arch Orthop Trauma Surg. 2023;143(7):4181-220.
3. Chidda A, Soares S, Tannast M, Schwab J, Seidel A. Long-term outcomes after a trimalleolar fracture involving the posterior malleolar fragment: an 11-year follow-up. Arch Orthop Trauma Surg. 2025;145(1):346.
4. Lambert LA, Stringer H, Weigelt L, Duncan L, Cowen J, Mason L. 2B or not 2B, should this not be the question? Comparison of 3D Surface Rendering CT to Plain Radiographs for Characterization of Posterior Malleolar Fracture Morphology. Foot Ankle Orthop. 2025;10(1):24730114241311879.
5. Mason LW, Marlow WJ, Widnall J, Molloy AP. Pathoanatomy and Associated Injuries of Posterior Malleolus Fracture of the Ankle. Foot Ankle Int. 2017;38(11):1229-35.
6. Haraguchi N, Haruyama H, Toga H, Kato F. Pathoanatomy of posterior malleolar fractures of the ankle. J Bone Joint Surg Am. 2006;88(5):1085-92. doi: 10.2106/JBJS.E.00856. Erratum in: J Bone Joint Surg Am. 2006;88(8):1835.
7. Bartoníček J, Rammelt S, Kostlivý K, Vaněček V, Klika D, Trešl I. Anatomy and classification of the posterior tibial fragment in ankle fractures. Arch Orthop Trauma Surg. 2015;135(4):505-16.
8. Audigé L, Bhandari M, Hanson B, Kellam J. A concept for the validation of fracture classifications. J Orthop Trauma. 2005;19(6):401-6.
9. Garbuz DS, Masri BA, Esdaile J, Duncan CP. Classification systems in orthopaedics. J Am Acad Orthop Surg. 2002;10(4):290-7.
10. Morales S, Massri-Pugin J, Mery P, Palma J, Filippi J, Villa A. Posterior Malleolar Fracture Assessment: An Independent Interobserver and Intraobserver Validation of Three Computed Tomography-Based Classifications. J Am Acad Orthop Surg Glob Res Rev. 2023;7(1):e22.00258.
11. Rashid MS, Islam R, Marsden S, Trompeter A, Teoh KH. Validation of three classification systems for posterior malleolus fractures of the ankle. Eur J Orthop Surg Traumatol. 2023;33(6):2601-8. Erratum in: Eur J Orthop Surg Traumatol. 2023;33(6):2609-10.
12. Kleinertz H, Mueller E, Tessarzyk M, Frosch KH, Schlickewei C. Computed tomography-based classifications of posterior malleolar fractures and their inter- and intraobserver reliability: a comparison of the Haraguchi, Bartoníček/Rammelt, and Mason classifications. Arch Orthop Trauma Surg. 2022;142(12):3895-902.
13. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106.
14. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.
15. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360-3.
16. Ma L, Mao J. Fisher Exact Scanning for Dependency. Journal of the American Statistical Association. 2018;114(525):245-58.
17. Cırdı YU, Demirel M, Kayaalp ME, Öztürk Rİ, Bozoğlan M, Salvi AG. Effect of clinician's experience and expertise on the inter- and intra-observer reliability of the computed tomography-based classification systems in posterior malleolus fractures. Acta Orthop Traumatol Turc. 2024;58(3):176-81.
18. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.
19. Keiler A, Riechelmann F, Thöni M, Brunner A, Ulmar B. Three-dimensional computed tomography reconstruction improves the reliability of tibial pilon fracture classification and preoperative surgical planning. Arch Orthop Trauma Surg. 2020;140(2):187-95.