rev bras ortop. 2018;53(1):2–9

SBOT
SOCIEDADE BRASILEIRA DE ORTOPEDIA E TRAUMATOLOGIA

RBO
REVISTA BRASILEIRA DE ORTOPEDIA

www.rbo.org.br

Original article

# Evaluation of interobserver agreement in Albertoni's classification for mallet finger☆

Vinícius Alexandre de Souza Almeida[a,*], Carlos Henrique Fernandes[a], João Baptista Gomes dos Santos[a], Francisco Alberto Schwarz-Fernandes[b], Flavio Faloppa[a], Walter Manna Albertoni[a]

[a] Departamento de Ortopedia e Traumatologia, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, SP, Brazil
[b] College of Medicine Orthopedic, University of South Florida, Tampa, United States

ARTICLE INFO

*Article history:*
Received 22 September 2016
Accepted 10 November 2016
Available online 12 December 2017

*Keywords:*
Tendon injuries
Finger injuries
Reproducibility of results
Classification
Rupture
Acquired hand deformities

ABSTRACT

*Objective:* To measure the reliability of Albertoni's classification for mallet finger.

*Methods:* Agreement study. Forty-three radiographs of patients with mallet finger were assessed by 19 responders (12 hand surgeons and seven residents). Injuries were classified by Albertoni's classification. For agreement comparison, lesions were grouped as: (A) tendon avulsion; (B) avulsion fracture; (C) fracture of the dorsal lip; and (D) physis injury–and subgroups (each group divided into two subgroups). Agreement was assessed by Fleiss's modification for kappa statistics.

*Results:* Agreement was excellent for Group A ($k = 0.95$ (0.93–0.97)) and remained good when separated into A1 and A2. Group B was moderate ($k = 0.42$ (0.39–0.44)) and poor when separated into B1 and B2. In the Group C, agreement was good ($k = 0.72$ (0.70–0.74)), but when separated into C1 and C2, it became moderate. Group D was always poor ($k = 0.16$ (0.14–0.19)). The general agreement was moderate, with ($k = 0.57$ (0.56–0.58)).

*Conclusion:* Albertoni's classification evaluated for interobserver agreement is considered a reproducible classification by the method used in the research.

☆ Study conducted at Disciplina de Cirurgia da Mão e Membro Superior, Departamento de Ortopedia e Traumatologia, Universidade Federal de São Paulo, São Paulo, SP, Brazil.

* *Corresponding author.*
E-mail: vinicius.asa@hotmail.com (V.A. Almeida).

https://doi.org/10.1016/j.rboe.2017.12.001

## Avaliação de concordância interobservador da classificação de Albertoni para dedo em martelo

*Palavras-chave:*
Traumatismos dos tendões
Traumatismos dos dedos
Reprodutibilidade dos testes
Classificação
Ruptura
Deformidades adquiridas da mão

RESUMO

*Objetivo:* Avaliar a reprodutibilidade da classificação de Albertoni para dedo em martelo.
*Métodos:* Foi feita uma avaliação por meio de questionário no qual foram avaliadas 43 radiografias em perfil da articulação interfalângica distal de dedos da mão, com lesão tipo dedo em martelo. Todas as lesões foram caracterizadas pela classificação de Albertoni, por 19 entrevistados (12 cirurgiões de mão e sete residentes). Foi então avaliada a concordância com o coeficiente Kappa generalizado, separadas por grupos – (A) avulsão tendínea; (B) fratura avulsão; (C) fratura do lábio dorsal e (D) lesão fisária – e por subgrupos (cada grupo dividido em 1 e 2).
*Resultados:* A concordância foi excelente para o grupo A (k = 0,95 [0,93-0,97]) e manteve-se boa quando separados em A1 e A2. No grupo B, a concordância foi moderada (k = 0,42 [0,39-0,44]), e foi ruim quando separada em B1 e B2. No grupo C, a concordância foi boa (k = 0,72 [0,70-0,74]), mas quando separada em C1 e C2 se tornou moderada. No grupo D foi sempre ruim (k = 0,16 [0,14-0,19]). A concordância geral foi moderada (k = 0,57 [0,56-0,58]).
*Conclusão:* Pela avaliação da concordância geral, a classificação de Albertoni é considerada reprodutível pelo método usado na pesquisa.

## Introduction

Lesions of the extensor mechanism of the fingers are among the most prevalent in the orthopedic practice. The terminal extensor tendon, formed by the union of two lateral slips, is inserted into the dorsal surface at the base of the distal phalanx. Injury of this tendon, or intra-articular fractures at the base of the distal phalanx, lead to a flexion deformity of the distal interphalangeal joint (DIPJ) known as mallet finger.[1] This lesion mainly affects the young population; it is common in sporting practices and may lead to a significant functional deficit if not treated properly.

Several clinical classifications have been described, aiming to categorize this condition. In 1957, Pratt et al.[2] classified mallet finger based on the etiology: laceration, crushing, and indirect trauma. In 1984, Wehbé and Schneider described a system that categorized these lesions into three types.[3] Doyle et al.[4] have also described another system widely used in the literature. In Brazil, Albertoni's.[5] clinical–radiological classification, described in 1986, is widely used.

A good quality classification should primarily be written in simple language and provide reliable guidelines to aid in treatment, prognosis, and reducing the possibility of complications. Moreover, it must be feasible, reliable, and reproducible; the latter characteristic is measured by interobserver agreement.[1,6] A classification is reproducible when several individuals are able to reproduce the same result at any time, anywhere.[1] Thus, it becomes possible to compare the results of different centers with different patients and the respective outcomes for each type of treatment.

Reproducibility studies are classic in the literature when measuring the quality of classification systems, especially in orthopedics. These studies usually include few observers, due to the difficulty in maintaining a reliable assessment. Any classification system worsens its agreement as the number of observers and categories increase. The low experience of observers in the assessed condition and multicenter studies also tend to decrease agreement.

No studies on the reproducibility of the Albertoni classification were retrieved in the literature, nor any study on the reproducibility of any mallet finger classification.

The authors conjectured that this classification has good interobserver agreement. This study is aimed at evaluating the interobserver agreement of the Albertoni classification for mallet finger, and to quantify its reproducibility in the management of this condition.

## Materials

This study was approved by the Research Ethics Committee of the institution where it was conducted (under CAAE No. 49960815.8.0000.5505).

A questionnaire survey was carried out in which 43 photographs of DIPJ radiographs in lateral view of hands with mallet finger injury were assessed. All radiographs were considered by the researchers to be of good quality.

The Albertoni classification was presented at the beginning of the questionnaire. It divides the lesions according to findings on a DIPJ radiograph in lateral view, categorizing them into four types: (A), pure tendon lesion without fracture; (B), bone avulsion lesion; (C), lesion associated with fracture of the dorsal region of the base of the distal phalanx, comprising one-third or more of the articular surface; and (D), epiphyseal detachment in children. Each type is divided into two subtypes. In types A and B, subtype 1 is characterized by a flexion deformity of less than 30° and subtype 2, by a flexion deformity greater than or equal to 30°. Deformities greater than 30° indicate injury to the retinacular ligaments and capsular structures in types A2 and B2. Type C is subdivided into C1, congruent joint (stable), and C2, sub-dislocated or dislocated joint (unstable). Type D is subdivided into D1, epiphyseal detachment (Salter and Harris lesion type 1) and D2, fracture-detachment (Salter and Harris type 3).[7,8]

Below each photograph of a radiograph, the options A1, A2, B1, B2, C1, C2, and D were presented, so that the observer could choose only one of them. Given the rarity of type D, it was not subdivided into D1 and D2. A goniometer and a pen were provided to the evaluator for the necessary measurements, so as to determine the subgroups 1 or 2 in type A and B, It was considered that all observers were able to measure the angles, as all have specialty degrees in orthopedics and traumatology.

The questionnaire was applied to 19 observers, all from the same institution, divided into 12 hand surgeons and seven hand surgery residents. Each observer answered the questionnaire separately, with no debate among them.

## Statistical methods

The interobserver agreement was calculated using the Fleiss-k agreement coefficient, a generalization for more than two evaluators based on Scott's agreement measure.[9,10] The standard error and consequently the confidence intervals were calculated according to Fleiss' algorithm.[11]

The agreements of the seven possible classification options were compared. The study also compared the agreement when the classification was grouped into four major categories: A1 and A2, in A; B1 and B2, in B; C1 and C2, in C; and D. In this way, it was possible to assess the difficulty of differentiating the groups (A, B, C, D) or the subgroups (1 and 2).

In all evaluations, in addition to the results from all 19 professionals, the results by the residents and hand surgeons were also compared separately. This comparison was made to assess the presence of differences of agreement between different levels of professional experience. For all inferential tests, the alpha error value was set at 0.05.

The *k*-value ranges from −1 to 1, where 1 means total agreement, −1 means total disagreement, and zero means that the evaluators classified the items at random. The agreement classification scale of Landis and Koch[12] was adopted (Table 1).

**Table 1 – Level of agreement, according to Landis and Koch's classification.**

| *k*-Value | Strength of agreement |
|---|---|
| <0.00 | No agreement |
| 0.01–0.20 | Very poor |
| 0.21–0.40 | Poor |
| 0.41–0.60 | Moderate |
| 0.61–0.80 | Good |
| 0.81–1.00 | Excellent |

**Table 2 – Number of responses for each Albertoni type, for each radiograph.**

| Radiographs | Options according to Albertoni classification | | | | | | |
|---|---|---|---|---|---|---|---|
| | A1 | A2 | B1 | B2 | C1 | C2 | D |
| 1 | 19 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 19 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 12 | 6 | 0 |
| 4 | 12 | 1 | 0 | 0 | 0 | 0 | 6 |
| 5 | 0 | 19 | 0 | 0 | 0 | 0 | 0 |
| 6 | 5 | 14 | 0 | 0 | 0 | 0 | 0 |
| 7 | 18 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 19 | 0 | 0 | 0 | 0 | 0 |
| 9 | 6 | 13 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 19 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 13 | 6 | 0 |
| 12 | 0 | 0 | 7 | 0 | 8 | 4 | 0 |
| 13 | 0 | 0 | 14 | 1 | 4 | 0 | 0 |
| 14 | 0 | 0 | 12 | 1 | 6 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 18 | 1 | 0 |
| 16 | 0 | 19 | 0 | 0 | 0 | 0 | 0 |
| 17 | 19 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 3 | 1 | 7 | 8 | 0 |
| 19 | 0 | 0 | 4 | 10 | 1 | 4 | 0 |
| 20 | 0 | 0 | 2 | 0 | 9 | 8 | 0 |
| 21 | 0 | 0 | 0 | 1 | 11 | 7 | 0 |
| 22 | 0 | 0 | 1 | 1 | 1 | 16 | 0 |
| 23 | 0 | 0 | 1 | 1 | 9 | 8 | 0 |
| 24 | 0 | 0 | 0 | 0 | 8 | 11 | 0 |
| 25 | 0 | 0 | 16 | 0 | 3 | 0 | 0 |
| 26 | 0 | 0 | 0 | 0 | 18 | 1 | 0 |
| 27 | 0 | 0 | 0 | 0 | 11 | 8 | 0 |
| 28 | 0 | 0 | 1 | 0 | 18 | 0 | 0 |
| 29 | 0 | 0 | 0 | 0 | 18 | 1 | 0 |
| 30 | 0 | 0 | 2 | 0 | 17 | 0 | 0 |
| 31 | 0 | 0 | 0 | 0 | 18 | 1 | 0 |
| 32 | 0 | 0 | 3 | 0 | 2 | 14 | 0 |
| 33 | 3 | 16 | 0 | 0 | 0 | 0 | 0 |
| 34 | 0 | 0 | 7 | 0 | 10 | 2 | 0 |
| 35 | 0 | 0 | 9 | 0 | 7 | 0 | 3 |
| 36 | 0 | 0 | 8 | 0 | 8 | 0 | 3 |
| 37 | 0 | 0 | 0 | 0 | 17 | 2 | 0 |
| 38 | 12 | 0 | 7 | 0 | 0 | 0 | 0 |
| 39 | 0 | 0 | 1 | 1 | 1 | 16 | 0 |
| 40 | 19 | 0 | 0 | 0 | 0 | 0 | 0 |
| 41 | 0 | 0 | 2 | 0 | 14 | 3 | 0 |
| 42 | 14 | 5 | 0 | 0 | 0 | 0 | 0 |
| 43 | 0 | 0 | 1 | 0 | 1 | 14 | 3 |

## Results

Table 2 presents the responses of the observers.

Regarding the distribution of types per observer group, the most recurrent type was C1 in the groups of surgeons and residents; the least recurrent type was D for surgeons and B2 for residents (Table 3).

When the Albertoni classification was grouped into larger categories for types A, B, C and D, the most prevalent type

**Table 3 – Distribution of the Albertoni classification types attributed by hand surgeons and residents to 43 images.**

| | The Albertoni classification | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Observer | A1 | | A2 | | B1 | | B2 | | C1 | | C2 | | D | |
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| *Surgeons* | | | | | | | | | | | | | | |
| 1 | 6 | 14.0 | 8 | 18.6 | 1 | 2.3 | 0 | 0.0 | 12 | 27.9 | 16 | 37.2 | 0 | 0.0 |
| 2 | 7 | 16.3 | 7 | 16.3 | 1 | 2.3 | 0 | 0.0 | 19 | 44.2 | 7 | 16.3 | 2 | 4.7 |
| 3 | 8 | 18.6 | 6 | 14.0 | 5 | 11.6 | 0 | 0.0 | 15 | 34.9 | 9 | 20.9 | 0 | 0.0 |
| 4 | 5 | 11.6 | 9 | 20.9 | 6 | 14.0 | 1 | 2.3 | 14 | 32.6 | 8 | 18.6 | 0 | 0.0 |
| 5 | 9 | 20.9 | 6 | 14.0 | 1 | 2.3 | 0 | 0.0 | 11 | 25.6 | 16 | 37.2 | 0 | 0.0 |
| 6 | 8 | 18.6 | 6 | 14.0 | 10 | 23.3 | 2 | 4.7 | 14 | 32.6 | 2 | 4.7 | 1 | 2.3 |
| 7 | 8 | 18.6 | 6 | 14.0 | 8 | 18.6 | 0 | 0.0 | 16 | 37.2 | 5 | 11.6 | 0 | 0.0 |
| 8 | 6 | 14.0 | 8 | 18.6 | 7 | 16.3 | 1 | 2.3 | 12 | 27.9 | 8 | 18.6 | 1 | 2.3 |
| 9 | 7 | 16.3 | 7 | 16.3 | 2 | 4.7 | 0 | 0.0 | 22 | 51.2 | 5 | 11.6 | 0 | 0.0 |
| 10 | 5 | 11.6 | 10 | 23.3 | 8 | 18.6 | 5 | 11.6 | 11 | 25.6 | 4 | 9.3 | 0 | 0.0 |
| 11 | 5 | 11.6 | 9 | 20.9 | 7 | 16.3 | 1 | 2.3 | 12 | 27.9 | 8 | 18.6 | 1 | 2.3 |
| 12 | 9 | 20.9 | 6 | 14.0 | 3 | 7.0 | 3 | 7.0 | 18 | 41.9 | 4 | 9.3 | 0 | 0.0 |
| **Total** | **83** | **16.1** | **88** | **17.1** | **59** | **11.4** | **13** | **2.5** | **176** | **34.1** | **92** | **17.8** | **5** | **1.0** |
| *Residents* | | | | | | | | | | | | | | |
| 13 | 7 | 16.3 | 7 | 16.3 | 12 | 27.9 | 0 | 0.0 | 11 | 25.6 | 5 | 11.6 | 1 | 2.3 |
| 14 | 7 | 16.3 | 8 | 18.6 | 2 | 4.7 | 0 | 0.0 | 13 | 30.2 | 11 | 25.6 | 2 | 4.7 |
| 15 | 6 | 14.0 | 9 | 20.9 | 4 | 9.3 | 1 | 2.3 | 11 | 25.6 | 12 | 27.9 | 0 | 0.0 |
| 16 | 6 | 14.0 | 9 | 20.9 | 7 | 16.3 | 1 | 2.3 | 15 | 34.9 | 4 | 9.3 | 1 | 2.3 |
| 17 | 5 | 11.6 | 8 | 18.6 | 5 | 11.6 | 1 | 2.3 | 12 | 27.9 | 10 | 23.3 | 2 | 4.7 |
| 18 | 6 | 14.0 | 8 | 18.6 | 7 | 16.3 | 0 | 0.0 | 17 | 39.5 | 3 | 7.0 | 2 | 4.7 |
| 19 | 7 | 16.3 | 8 | 18.6 | 6 | 14.0 | 1 | 2.3 | 15 | 34.9 | 4 | 9.3 | 2 | 4.7 |
| **Total** | **44** | **14.6** | **57** | **18.9** | **43** | **14.4** | **4** | **1.3** | **94** | **31.2** | **49** | **16.3** | **10** | **3.3** |
| **Grand total** | **127** | **15.5** | **145** | **17.7** | **102** | **12.6** | **17** | **2.1** | **270** | **33.0** | **141** | **17.3** | **15** | **1.8** |

**Table 4 – Distribution of the Albertoni classification types, grouped into larger categories, attributed by hand surgeons and residents to 43 images.**

| | Grouped Albertoni classification | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | | B | | C | | D | |
| | *n* | % | *n* | % | *n* | % | *n* | % |
| *Surgeons* | | | | | | | | |
| | 14 | 32.6 | 1 | 2.3 | 28 | 65.1 | – | – |
| | 14 | 32.6 | 1 | 2.3 | 26 | 60.5 | 2 | 4.7 |
| | 14 | 32.6 | 5 | 11.6 | 24 | 55.8 | – | – |
| | 14 | 32.6 | 7 | 16.3 | 22 | 51.2 | – | – |
| | 15 | 34.9 | 1 | 2.3 | 27 | 62.8 | – | – |
| | 14 | 32.6 | 12 | 27.9 | 16 | 37.2 | 1 | 2.3 |
| | 14 | 32.6 | 8 | 18.6 | 21 | 48.8 | – | – |
| | 14 | 32.6 | 8 | 18.6 | 20 | 46.5 | 1 | 2.3 |
| | 14 | 32.6 | 2 | 4.7 | 27 | 62.8 | – | – |
| | 15 | 34.9 | 13 | 30.2 | 15 | 34.9 | – | – |
| | 14 | 32.6 | 8 | 18.6 | 20 | 46.5 | 1 | 2.3 |
| | 15 | 34.9 | 6 | 14.0 | 22 | 51.2 | – | – |
| **Total** | **171** | **33.1** | **72** | **14.0** | **268** | **51.9%** | **5** | **1.0** |
| *Residents* | | | | | | | | |
| | 14 | 32.6 | 12 | 27.9 | 16 | 37.2 | 1 | 2.3 |
| | 15 | 34.9 | 2 | 4.7 | 24 | 55.8 | 2 | 4.7 |
| | 15 | 34.9 | 5 | 11.6 | 23 | 53.5 | – | – |
| | 15 | 34.9 | 8 | 18.6 | 19 | 44.2 | 1 | 2.3 |
| | 13 | 30.2 | 6 | 14.0 | 22 | 51.2 | 2 | 4.7 |
| | 14 | 32.6 | 7 | 16.3 | 20 | 46.5 | 2 | 4.7 |
| | 15 | 34.9 | 7 | 16.3 | 19 | 44.2 | 2 | 4.7 |
| **Total** | **101** | **33.6** | **47** | **15.6** | **143** | **47.5** | **10** | **3.3** |
| **Grand total** | **272** | **33.3** | **119** | **14.6** | **411** | **50.3** | **15** | **1.8** |

**Table 5 – General agreement coefficient for the evaluation of images with *p*-value < 0.0001.**

| Classification | *k* (95% confidence interval) | Classification | *k* (95% confidence interval) |
|---|---|---|---|
| *Surgeon* | | | |
| A | 0.95 (0.91–0.99) | A1 | 0.75 (0.71–0.78) |
| | | A2 | 0.84 (0.80–0.87) |
| B | 0.34 (0.31–0.38) | B1 | 0.34 (0.30–0.37) |
| | | B2 | 0.19 (0.15–0.23) |
| C | 0.71 (0.67–0.75) | C1 | 0.51 (0.48–0.55) |
| | | C2 | 0.44 (0.41–0.48) |
| D | 0.10 (0.06–0.14) | D | 0.10 (0.06–0.14) |
| General | 0.72 (0.69–0.74) | | 0.56 (0.54–0.58) |
| *Resident* | | | |
| A | 0.96 (0.89–1.00) | A1 | 0.82 (0.76–0.89) |
| | | A2 | 0.90 (0.83–0.96) |
| B | 0.55 (0.48–0.61) | B1 | 0.46 (0.39–0.52) |
| | | B2 | 0.49 (0.43–0.56) |
| C | 0.77 (0.70–0.83) | C1 | 0.49 (0.43–0.56) |
| | | C2 | 0.37 (0.30–0.43) |
| D | 0.24 (0.18–0.31) | D | 0.24 (0.18–0.31) |
| General | 0.76 (0.72–0.81) | | 0.59 (0.55–0.62) |
| *Surgeon + Resident* | | | |
| A | 0.95 (0.93–0.97) | A1 | 0.77 (0.74–0.79) |
| | | A2 | 0.86 (0.84–0.88) |
| B | 0.42 (0.39–0.44) | B1 | 0.38 (0.36–0.40) |
| | | B2 | 0.28 (0.26–0.30) |
| C | 0.72 (0.70–0.74) | C1 | 0.52 (0.50–0.54) |
| | | C2 | 0.42 (0.39–0.44) |
| D | 0.16 (0.14–0.19) | D | 0.16 (0.14–0.19) |
| General | 0.73 (0.71–0.74) | | 0.57 (0.56–0.58) |

was C, followed by A, B, and D, both in the hand surgeons and residents groups (Table 4).

Table 5 presents the results for the agreement in the Albertoni classification according to the Fleiss agreement coefficient.

Among the hand surgeons, classifications A1 ($k = 0.75$ [0.71–0.78]) and A2 ($k = 0.84$ [0.80–0.87]) presented higher agreement than the other groups, having good and excellent agreements, respectively. In types B1 ($k = 0.34$ [0.30–0.37]) and B2 ($k = 0.19$ [0.15–0.23]), the agreement was poor and very poor, respectively. Types C1 ($k = 0.51$ [0.48–0.55]) and C2 ($k = 0.44$ [0.41–0.48]) had moderate agreement, whereas in type D ($k = 0.10$ [0.06–0.14]), the agreement was very poor. The general agreement presented $k = 0.56$ (0.54–0.58), was considered moderate.

Among the residents, the types A1 ($k = 0.82$ [0.76–0.89]) and A2 ($k = 0.90$ [0.83–0.96]) also presented better agreement than the others, and were considered to be excellent. The groups B1 ($k = 0.46$ [0.39–0.52]), B2 ($k = 0.49$ [0.43–0.56]), and C1 ($k = 0.49$ [0.43–0.56]) presented moderate agreement, while groups C2 ($k = 0.37$ ([0.30–0.43]) and D ($k = 0.24$ [0.18–0.31]), presented poor agreement. However, taking into account the confidence interval, it can be considered that groups B1 and B2 and C1 and C2 had the same agreement. The general agreement was $k = 0.59$ (0.55–0.62), considered moderate.

When hand surgeons and residents were analyzed together, the results were similar. In A1 ($k = 0.77$ [0.74–0.79]) and A2 ($k = 0.86$ [0.84–0.88]), the agreement was considered good and excellent, respectively. In turn, in B1 ($k = 0.38$ [0.36–0.40]) and B2 ($k = 0.28$ [0.26–0.30]) the agreements were considered poor. Types C1 ($k = 0.52$ [0.50–0.54]) and C2 ($k = 0.42$ [0.39–0.44]) presented moderate agreement and type D ($k = 0.16$ [0.14–0.19]), very poor agreement. The general agreement was $k = 0.57$ (0.56–0.58), considered moderate.

When assessing the classification into large groups (A, B, C, D), an improvement was observed for all groups, and the most significant improvement was observed in group C. Taking into account the 95% confidence interval, in all analyses combining subtypes 1 and 2 improved the agreement.

Among the surgeons, the agreement in A ($k = 0.95$ [0.91–0.99]) was excellent, in B it was poor ($k = 0.34$ [0.31–0.38]), in C ($k = 0.71$ [0.67–0.75]) it was good and, in D it remained very poor. The overall agreement was good, with $k = 0.72$ (0.69–0.74).

Among the residents, the agreement in A ($k = 0.96$ [0.89–1.00]) was excellent, in B ($k = 0.55$ [0.48–0.61]), moderate, in C ($k = 0.77$ [0.70–0.83]), good, and in D it remained poor. The overall agreement was good, with $k = 0.76$ (0.72–0.81).

When combining hand surgeons and residents, the agreement in A ($k = 0.95$ [0.93–0.97]) was excellent, in B ($k = 0.42$ [0.39–0.44]), moderate, in C ($k = 0.72$ [0.70–0.74]), good, and in D it remained poor. The general agreement in this second evaluation was good, with $k = 0.73$ (0.71–0.74). The data plotted in the graphs present the respective confidence intervals (Figs. 1 and 2).

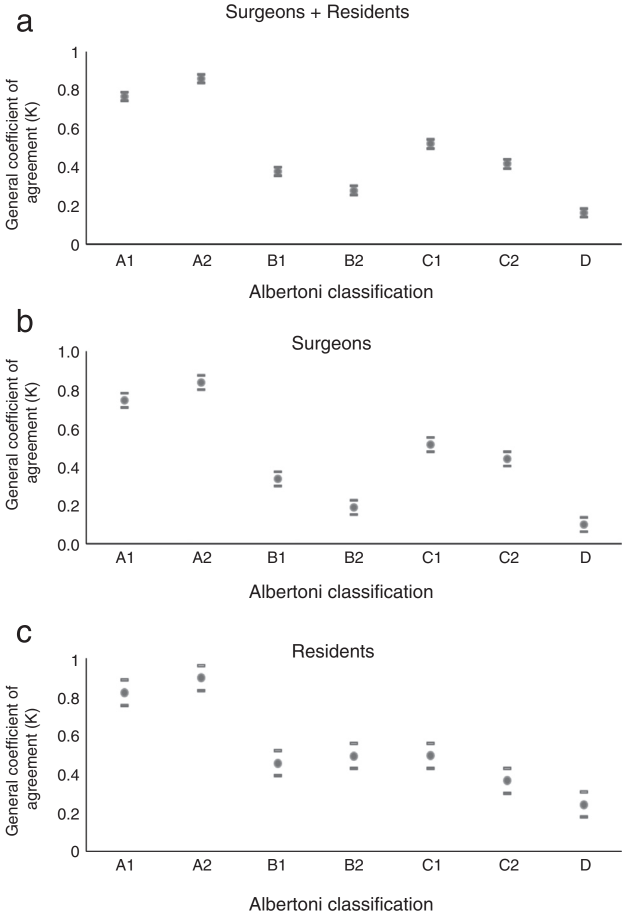

**Fig. 1 – Generalized coefficient of agreement for each subgroup. The dots represent the coefficient value and the dashes, the 95% confidence interval. (a) Surgeons and residents, (b) hand surgeons, and (c) hand surgery residents.**

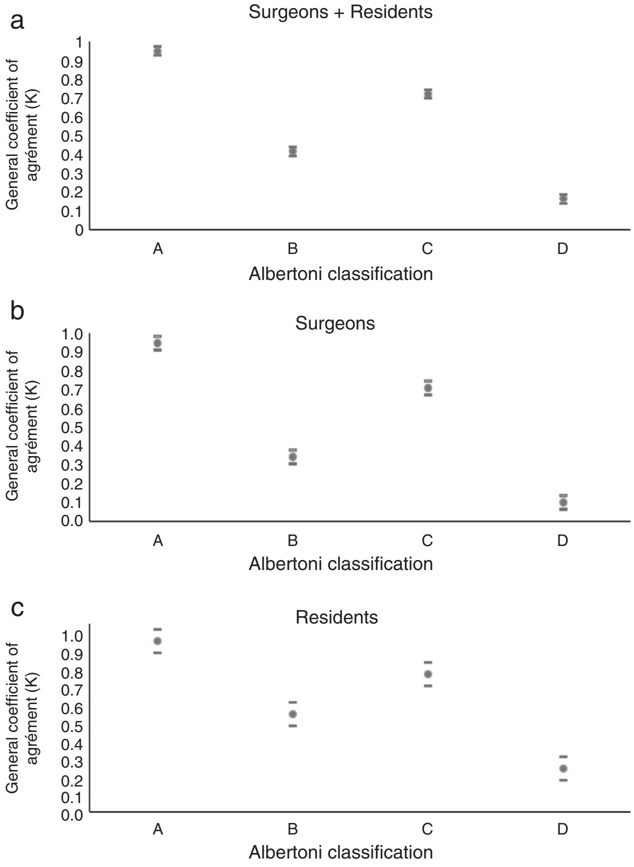

**Fig. 2 – Generalized coefficient of agreement for the grouped data. The dots represent the coefficient value and the dashes, the 95% confidence interval. (a) Surgeons and residents, (b) hand surgeons, and (c) hand surgery residents.**

## Discussion

Mallet finger is a very prevalent condition, mainly affecting an economically active age population. An efficient and reproducible classification is able to guide the professional to a more efficient treatment.

Among the previously described classifications for this type of deformity, that of Pratt et al.[2] only divides the etiology and does not lead to treatment or prognosis. The classification of Wehbé and Schneider,[3] which uses the extent of joint surface involvement, has the same limitation. It stratifies each of the three types into three subtypes (A, less than one-third of the affected joint surface, B between one-third and two-thirds of the articular surface, and C, more than two-thirds of the affected surface). However, this does not provide guidance regarding treatment or prognosis. Doyle's classification provides a stratification based on clinical parameters that remain uncontemplated in the other systems, but there is no categorization of the radiographic patterns for all types.[4] The latter is the most widely used in literature worldwide.[1]

Compared with other classifications, the Albertoni classification defines a therapeutic schedule, which changes according to the lesion in question. Each type has a specific treatment: type A1 and B1 injuries are classically treated with immobilization. Types C1 and D are treated with non-surgical reduction and immobilization with a metal splint. Type A2, B2, and C2 lesions usually require surgical treatment.[5,8] This classification is widely used among Brazilian orthopedic surgeons, hence the importance of assessing its reproducibility.

In the literature, no study on the reproducibility of the Albertoni classification was retrieved, nor was any study on the reproducibility of mallet finger classifications. This demonstrates the originality of the present study.

The Albertoni classification is based on radiographs. In the literature, several interobserver agreement studies have evaluated radiographic classifications. Audigé et al.[13] evaluated 44 reproducibility studies on orthopedic classifications with the use of imaging criteria and found little uniformity in the methodology, which hinders the comparison of reproducibility. Belloti et al.[6] assessed the reproducibility of distal radius fracture classifications, while Utino et al.[14] studied the agreement in the AO classification for long bones in the pediatric population.

The number of radiographs under evaluation is an important factor for assessing agreement. Both too few and too many evaluations tend to worsen agreement.[15] In Audigé's systematic review of orthopedic reproducibility studies, a large variation in the number of radiographs per study (from 14 to 200 evaluations) was observed.[13] Berger et al.,[16] in a systematic review on the reproducibility of the Eaton classification for rhizarthrosis, assessed four studies. In these studies, the number of radiographs ranged from 40 to 43. Based on these studies, the authors consider that the 43 radiographs used in the present study were sufficient to evaluate the interobserver agreement in the Albertoni classification.

The number of observers is another factor that interferes with the coefficient of agreement. The greater the quantity, the lower the probability of agreement. The literature also did not present uniformity regarding the number of observers. Thomsen et al.[17] used only four observers, while Randsborg and Sivertsen[18] worked with 12 observers and Audigé et al.[13] analyzed works with 2–36 observers (median = 5). The present study included 19 observers in subgroups of residents and hand surgeons. The number of radiographs and observers in this study is therefore within the norms found in the literature.[6,13,14,16–24]

The higher the number of categories in a classification, the worse the agreement.[6,18] Albertoni's classification, with seven possible options, would tend to have a worse agreement when compared with other classifications. This was demonstrated when grouping types A1 and A2 into A, B1 and B2 into B, and C1 and C2 into C; an increase of agreement was observed in all the analyses, since the number of categories had decreased to four.

There is no consensus in the literature regarding which is the cut-off value of $k$ to consider a classification as reproducible.[13,15] These values are arbitrarily defined by the authors.[15] Fleiss[9] consider $k$-values between 0.40 and 0.75 to present moderate to good agreement. Svanholm et al.[25] only consider as good values of $k$ greater than 0.75. In turn, Brage et al.[22] consider $k$-values above 0.50 as reproducible. Landis and Koch,[12] the parameter used in the present study (Table 1) and the most used today,[13] considered moderate agreement as those in the range of 0.4–0.6, and good agreement, above 0.6.

When assessed as a single group, Albertoni types A1 and A2 presented $k = 0.95$ (0.93–0.97), which indicates an excellent agreement. When stratified into A1 ($k = 0.77$ [0.74–0.79]) and A2 ($k = 0.84$ [0.80–0.87]), the coefficient of agreement decreased slightly, but remained good. The $k$ coefficient presented a decrease of 16%, *i.e.*, there is little change in the agreement when combining the categories and removing the parameter that differentiates them. This shows that an angle of <30° (A1) or >30° (A2) can be considered a reproducible parameter.

Albertoni types B1 and B2, when evaluated as a single group, presented a moderate agreement coefficient, with ($k = 0.42$ [0.39–0.44]). When stratified into B1 ($k = 0.38$ [0.36–0.40]) and B2 ($k = 0.28$ [0.26–0.30]), the coefficient of agreement presented a 21% decrease, being classified as poor. Similarly to the findings for type A, there is no difficulty in differentiating between B1 and B2, as well as between A1 and A2, which corroborates the fact that the angle parameter is reproducible. Regarding the poor agreement, the authors believe that the main reason for this result is the low prevalence of type B in the questionnaire applied (14.6%), as seen in Table 4.

When evaluated as a single group, Albertoni types C1 and C2 presented a good agreement coefficient, with $k = 0.77$ (0.70–0.83). When stratified into C1 ($k = 0.52$ [0.50–0.54]) and C2 ($k = 0.42$ [0.39–0.44]), the coefficient of agreement presented a pronounced decrease of 41%, which became moderate for C1 and poor for C2. This leads to the assumption that the joint congruence of the DIPJ is difficult to define. That is, in relation to joint congruence, agreement decreased considerably. The authors believe that one of the ways to improve agreement in type C would be to better define the criterion for joint congruence. This may guide future modifications to this classification.

Another relevant factor is that out of the 43 radiographs, in 22 the observers presented doubt in the choice between B and C (Table 2). In contrast, when type A (A1 or A2) was chosen, it was always concordant, except for radiographs 4 and 38 (Table 2). This was evidenced by the lower agreement in types B and C when compared with type A. The authors believe that the parameter to distinguish between bone avulsion (type B) and fracture of the dorsal region of the base of the distal phalanx (type C) is not well understood by the observers.

Due to the low incidence of this type of lesion, type D was not separated in D1 and D2. A very poor agreement was observed, with $k = 0.16$ (0.14–0.19). The low prevalence of this type (1.8%), in only four cases, also justifies the low agreement.[23]

The evaluators' experience in the assessed condition tends to change agreement.[18] Mattos et al.[24] concluded that the lack of experience of the observers decreased agreement. However, in the present study, among the residents, the general agreement was $k = 0.76$ (0.72–0.81), while among hand surgeons it was $k = 0.72$ (0.69–0.74). Although the group of residents presented $k$-values greater than the group of hand surgeons, considering the 95% confidence interval (Figs. 1 and 2), it cannot be stated that the agreement is higher in that group than in the group of surgeons. This is an advantage of the Albertoni classification, in which the agreement is not altered with the experience of the observers. The authors believe that there is a tendency for greater agreement among residents because they have similar levels of knowledge and experience, and because they are in training in the same center, which leads to uniformity. However, this was not demonstrated in the present study.

Audigé mentions that in the evaluation of 44 studies on the reproducibility of orthopedic classifications, of the 86 coefficients of agreement calculated, only four were excellent ($k > 0.80$), 17 were good (between 0.60 and 0.80), 32 were moderate (0.40–0.60), and 33 were fair or poor (<0.40).[13] The overall agreement in the Albertoni classification was moderate, with $k = 0.57$ (0.56–0.58), based on the classification by Landis and Koch.[12] Despite the moderate agreement, when compared with the literature and taking into account all the factors discussed, the authors consider the Albertoni classification to be reproducible.

A selection bias has to be considered, as the radiographs in the present study were not randomized, having been chosen by the researchers for their quality. Another relevant feature is that it cannot be guaranteed that measurements were made using the appropriate methods, even though observers were correctly instructed and provided correct measurement material. This was a single-center study, which tends to homogenize responses and improve agreement.

## Conclusion

The Albertoni classification presented good or excellent interobserver agreement for types A1 and A2, moderate for types C1 and C2, and poor for types B1, B2, and D. Following the statistical methods employed, and compared with research in the literature, the authors consider the Albertoni classification to be reproducible. The authors believe that a better definition of the criteria for joint congruence would substantially improve agreement.

## Conflicts of interest

The authors declare no conflicts of interest.

## REFERENCES

1. Alla SR, Deal ND, Dempsey IJ. Current concepts: mallet finger. Hand. 2014;9(2):138–44.
2. Pratt DR, Bunnell S, Howard LD. Mallet finger: classification and methods of treatment. Am J Surg. 1957;93(4):573–9.
3. Wehbé MA, Schneider LH. Mallet fractures. J Bone Joint Surg Am. 1984;66(5):658–69.
4. Doyle JR, Green DP, Hotchkiss RN, Pederson WC. Extensor tendons: acute injuries. In: Green DP, Hotchkiss RN, Pederson WC, editors. Green's operative hand surgery. 4th ed. New York: Churchill Livingstone; 1999. p. 195–8.
5. Albertoni WM. Estudo crítico de tratamento do dedo em martelo. Análise de 200 caso [tese]. São Paulo: Universidade Federal de São Paulo, Escola Paulista de Medicina; 1986.
6. Belloti JC, Tamaoki MJ, Franciozi CE, Santos JB, Balbachevsky D, Chap EC, et al. Are distal radius fracture classifications reproducible? Intra and interobserver agreement. Sao Paulo Med J. 2008;126(3):180–5.
7. Albertoni WM. Mallet finger: classification. Rev Hosp Sao Paulo Esc Paul Med. 1989;1(3):133–6.
8. Albertoni WM. The Brooks-Graner procedure for correction of mallet finger. Hand. 1988;3:97–100.
9. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82.
10. Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q. 1955;19(3):321–5.
11. Fleiss JL, Nee JC, Landis JR. Large sample variance of kappa in the case of different sets of raters. Psycho Bull. 1979;86(5):974–7.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
13. Audigé L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. Acta Orthop. 2004;75(2):184–94.
14. Utino AY, de Alencar DR, Maringolo LF, Negrão JM, Blumetti FC, Dobashi ET. Concordância intra e interobservadores do sistema de classificação AO para fraturas dos ossos longos na população pediátrica. Rev Bras Ortop. 2015;50(5): 501–8.
15. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257–68.
16. Berger AJ, Momeni A, Ladd AL. Intra and interobserver reliability of the Eaton classification for trapeziometacarpal arthritis: a systematic review. Clin Orthop Relat Res. 2014;472(4):1155–9.
17. Thomsen NO, Overgaard S, Olsen LH, Hansen H, Nielsen ST. Observer variation in the radiographic classification of ankle fractures. Bone Joint J. 1991;73(4):676–8.
18. Randsborg PH, Sivertsen EA. Classification of distal radius fractures in children: good inter-and intraobserver reliability, which improves with clinical experience. BMC Musculoskelet Disord. 2012;13:6.
19. Van Embden D, Rhemrev SJ, Genelin F, Meylaerts SAG, Roukema GR. The reliability of a simplified Garden classification for intracapsular hip fractures. Orthop Traumatol Surg Res. 2012;98(4):405–8.
20. Valderrama-Molina CO, Estrada-Castrillón M, Hincapie JA, Lugo-Agudelo LH. Intra and interobserver agreement on the Oestern and Tscherne classification of soft tissue injury in periarticular lower-limb closed fractures. Colomb Med (Cali). 2014;45(4):173–8.
21. Kim JK, Kim DJ. The risk factors associated with subluxation of the distal interphalangeal joint in mallet fracture. J Hand Surg Eur Vol. 2015;40(1):63–7.
22. Brage ME, Rockett M, Vraney R, Anderson R, Toledano A. Ankle fracture classification: a comparison of reliability of three X-ray views versus two. Foot Ankle Int. 1998;19(8):555–62.
23. Liggieri AC, Tamanaha MJ, Abechain JJK, Ikeda TM, Dobashi ET. Concordância intra e interobservadores das diferentes classificações usadas na doença de Legg-Calvé-Perthes. Rev Bras Ortop. 2015;50(6):680–5.

24. Mattos CA, Jesus AAK, dos Santos Floter M, Nunes LFB, de Baptista Sanches B, Zabeu JLA. Reprodutibilidade das classificações de Tronzo e AO para fraturas transtrocanterianas. Rev Bras Ortop. 2015;50(5):495–500.

25. Svanholm H, Starklint H, Gundersen HJ, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS. 1989;97(8):689–98.