Assessment of the reliability and reproducibility of the Langenskiöld classification in Blount's disease

Jason du Plessis[a], Gregory B. Firth[b], and Anthony Robertson[c]

The Langenskiöld classification is the most commonly utilized classification system for the radiological features of Blount's disease. Although there is only a single study found on the interobserver variability and none found on the intraobserver variability, it is commonly used for prognostication and guiding management decisions. The aim of this study was to determine the reliability and reproducibility of the Langenskiöld classification. A retrospective review of radiographs was done of patients treated for infantile and juvenile Blount's disease at Chris Hani Baragwanath Academic Hospital from 2006 to 2016. There were 70 radiographs of acceptable quality, which were reviewed and staged on two occasions according to the Langenskiöld classification by three orthopaedic consultants and three orthopaedic surgery senior residents. Pearson correlation coefficients, percentage agreements, and κ statistics were used to evaluate both the reliability and reproducibility. Of the 70 images staged, only two (2.9%) were staged the same by all six observers, and 20 (28.6%) images differed by a single stage. The consultants had 17 (24.3%) images staged the same whereas the residents had 12 (17.1%) images staged the same. The overall κ for all six observers showed a fair agreement of 0.24. Again, the consultants had a higher κ-value compared to residents of 0.25 and 0.24, respectively. The reproducibility amongst all observers was fair with a κ-value of 0.38. The consultants had a higher mean score of 0.48 compared to 0.26 for the residents. There was only a fair overall reliability and reproducibility amongst the six observers. We recommend the Langenskiöld classification be used with caution when being used for prognostication and management planning as well as when interpreting any research relying on this classification. Level of evidence: Level III, diagnostic study. *J Pediatr Orthop B* **29:311–316** **Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.**

Journal of Pediatric Orthopaedics B 2020, **29**:311–316

Keywords: Blount's disease, classification, interobserver, intraobserver, Langenskiold, reliability, reproducibility, tibia vara

[a]University of the Witwatersrand, Johannesburg, [b]Chris Hani Baragwanath Academic, Hospital, Soweto and [c]Charlotte Maxeke Johannesburg, Academic Hospital, Johannesburg, South Africa

Correspondence to Jason du Plessis, MBChB, University of the Witwatersrand, 7 York Rd, Parktown, Johannesburg, 2193, South Africa
Tel: +27 82 445 7247; e-mail: jasondp_29@yahoo.com

Downloaded from http://journals.lww.com/jpo-b by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdtwnfKZBYtws= on 10/11/2021

Introduction

Blount's disease was first described in 1922 by Phillip Erlacher in a case report [1]. The condition was then later divided into three groups: infantile, juvenile and adolescent depending on the age at presentation [1]. The Langenskiöld classification was introduced in 1952 after a review of 23 cases of infantile Blount's disease, where six radiographic stages were distinguished as a progression of the disease [2]. Langenskiöld described the evolution of infantile Blount's disease through six radiographic changes with increasing maturation of the skeleton (Table 1). Traditionally, Blount's disease has been classified using this six-stage system to represent the spectrum of disease severity and is used for both infantile and juvenile Blount's disease [3].

Table 1 Description of the Langenskiöld stages [2]

Stage I	Medial metaphyseal beaking
Stage II	Saucer-shaped defect in the medial metaphysis
Stage III	Saucer deepens into a step
Stage IV	Epiphysis bent over the medial beak
Stage V	Double epiphysis
Stage VI	Medial bony bar evident

The Langenskiöld classification remains the most commonly used classification in Blount's disease despite the observation of frequent recurrence in cases treated in stages I–III with osteotomy, which would normally not be expected [4]. Only a single study by Stricker *et al.* [5] could be located investigating the interobserver reliability, with no studies recorded investigating the intraobserver reliability. The poor interobserver variability in that study was attributed by Langenskiöld to the gradual progression from one stage to another [3].

Multiple authors have examined the Langenskiöld classification's usefulness regarding the prognostication of spontaneous resolution as well as recurrence. In a series of 22 Stage I limbs, Shinohara *et al.* [6] recommended observation for children staged up to a Langenskiöld stage III based on their levels of resolution, which was correlated to Langenskiöld stage. Laville *et al.* [7] concluded that despite Langenskiöld Stages II or III, certain cases of Blount's disease may resolve spontaneously without any intervention and may account for some of the positive results reported in bracing studies. Khanfour [8] found recurrence in his patients treated with gradual correction using the Ilizarov method, with the recurrence mostly

1060-152X Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved. DOI: 10.1097/BPB.0000000000000692

Copyright © 2020 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

found in children with a higher initial Langenskiöld stage. The Langenskiöld classification was found to be the only independent factor affecting the prognosis in late onset tibia vara, as well as being found to correlate with the long-term persistence of symptoms after skeletal maturity following treatment [9]. The medial physeal slope angle is another more objective radiographic parameter looked into by Kling *et al*., and was found to be an independent predictor of recurrence following osteotomy [10].

Numerous authors have also recommended treatment based on the Langenskiöld classification system. In the early stages, some authors have advocated the use of antivarus bracing [11], although its effectiveness still remains uncertain in light of the natural history and possible spontaneous resolution [1]. Loder *et al*. reported that there were worsening results of correction and increased deformity recurrence when surgery was performed at higher Langenskiöld stages [12], whereas knees corrected before Stage III were found to have a better outcome [4,13]. In light of these observations, surgical correction has been advocated for patients staged II or III after age three and from stage IV at any age [9].

The aim of this study was to assess the usefulness of the Langenskiöld classification in clinical practice and decision making by assessing its reliability and reproducibility.

Materials and methods

This was a retrospective record review of radiographs taken from patients treated at a tertiary Academic Hospital for Blount's disease from 2006 to 2016.

Ethics Approval was obtained from the University's Human Research Ethics Committee (HREC) (Medical) before data collection.

The preoperative radiographs were identified from all patients with Blount's disease on an electronic database. Some of the radiographs were immediately preoperative, while others dated back to the first presentation of the patient long before surgery. This allowed for a range of images in terms of severity and deformity. All patients in the database with adequate radiographs showing and antero-posterior (AP) views of the affected knees were included. The radiographs were a combination of standing AP, supine AP and standing long leg views that we used for preoperative assessment. These radiographs had to be adequately exposed to show sufficient bony detail to allow for staging. All radiographs where the exposure was inappropriate or where the bony detail was obscured by artefact or poor patient positioning were excluded.

Data collection

A retrospective review of patient radiographs was conducted from a database of patients who were treated for infantile and juvenile Blount's disease at a tertiary academic hospital between 2006 and 2016. A total of 72 patients including 95 knees were evaluated in the records for good quality AP knee radiographs. All patient names were removed when the images were presented but the ages were clearly visible, and each image was saved with a unique number.

Six observers were selected to participate in the study from the University's Department of Orthopaedic Surgery. Three paediatric orthopaedic surgery consultants and three senior residents currently rotating through the Paediatric Orthopaedic Units at two tertiary academic teaching hospital affiliated to the University of Witwatersrand participated in the study.

All qualifying radiographs were combined with the original image of the Langenskiöld classification as well as a worded description taken from the original article [2] to create the survey image (Fig. 1) [14]. This was done individually for all 72 images to eliminate the need for any recall or memory of the classification. An online survey was created on www.surveymonkey.com (© 2017 SurveyMonkey Inc, San Mateo, California, USA) with images of qualifying radiographs together with the classification system and a worded description of each stage. All observers were brought together in a single session where presentations on the details of the Langenskiöld classification were reviewed and all reviewers were taken through the classification along with the worded description. They were then all briefed on the aim and objectives of the study as well as how these would be achieved. An example survey was displayed in a presentation to all observers and all observers were familiarized with SurveyMonkey and how the surveys would work.

The first survey was then compiled with all included images, as well as asking for the Rank of the participant (consultant or resident) and their professional registration number. This allowed for easier anonymous cross comparison between ranks over the two surveys. An access link to the survey was created and emailed to all of the observers who then followed the link and completed the survey on their computer or smartphones by staging each radiograph according to the Langenskiöld classification. The radiographs were then reordered for the second survey using Microsoft Excel 2013 to randomly generate new numbers. The new order of images was recorded on a spreadsheet and a new survey was created. Two weeks after the last person completed the first survey, the link to the newly created second survey was sent to all observers to restage each image again. This second survey allowed for the evaluation of both the reliability and the reproducibility of the classification system between all observers as well as to evaluate the consultants separately from the residents.

Data analysis

The data obtained from each survey was tabulated using Microsoft Excel 2013. All the observers were

Copyright © 2020 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Fig. 1

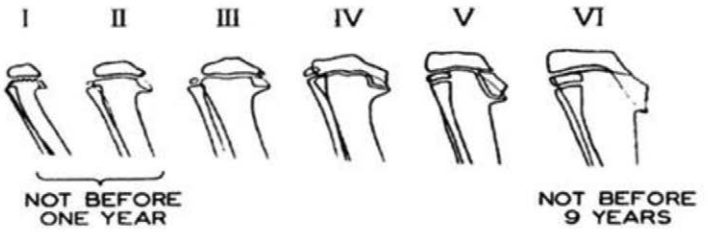


Figure 1: Diagram of the roentgenographic changes seen in the infantile type of tibia vara and their development with increasing age. (from Langenskiöld, A. and E.B. Riska, *Tibia vara (osteochondrosis deformans tibiae).* J Bone Joint Surg Am, 1964. **46**(7): p. 1405-1420

Table 1: Description of the Langenskiöld stages (4)
Stage I: Medial metaphyseal beaking
Stage II: Saucer shaped defect in the medial metaphysis
Stage III: Saucer deepens into a step
Stage IV: Epiphysis bent over the medial beak
Stage V: Double epiphysis
Stage VI: Medial bony bar evident

Example of image provided for staging. Images on left reproduced with permission [14].

allocated a number based on their unique professional number provided, allowing the results from the same observer to be allocated to the correct slot in the table. The answer given for each image in terms of the Langenskiöld staging was recorded in this table for each observer.

The data analysis was initially performed using Stata/IC version 14.1. The Cohens Kappa statistic was initially used to obtain κ-values for the reliability and reproducibility between observers using the mean values from the datasheet. Following that, the Fleiss Kappa statistic was calculated using Microsoft Excel 2013 to obtain combined κ-values for all observers, and separately to compare consultants and residents.

Additional analysis was performed using Graphpad Prism version 5.02 (Graphpad.com). Here the spread and distribution of results was analyzed. The χ^2 analysis was also performed to assess the significance of the relationship between categorical values, for which a P-value <0.05 was considered statistically significant. For the purposes of this study, the number of images staged as V or VI was combined due to the small numbers in these groups.

Results

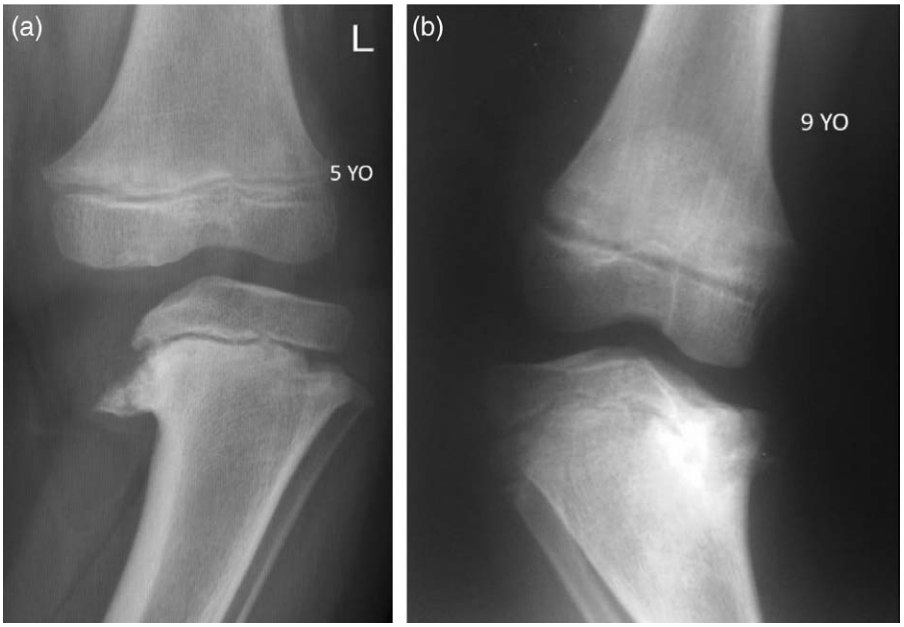
There were 70 images created for review from 62 patients who were found to have acceptable images for staging. There were 25 images excluded on the basis of poor rotation, poor exposure or overlying markings affecting visibility of the bony detail. Patients ranged from 2 to 9 years of age and were all diagnosed with and subsequently treated for Blount's disease. All the images chosen showed adequate bony detail in the proximal tibia to allow for visualization of the changes associated with each stage.

Reliability

Survey 1 was primarily used to calculate the reliability of the Langenskiöld classification based on the responses of the six observers. There were only two images (2.9%) shown in Fig. 2 that were classified the same by all six observers. There were 20 images where all the observers differed by only a single stage (28.6%), which means there were only 22 images (31.5%), where the six observers either staged an image the same or differed by a single stage. A similar analysis was done on the second survey, which showed a slight improvement with 27 images (38.6%) being classified either the same or only differing by a single stage. Over the two surveys, the most common pairing where

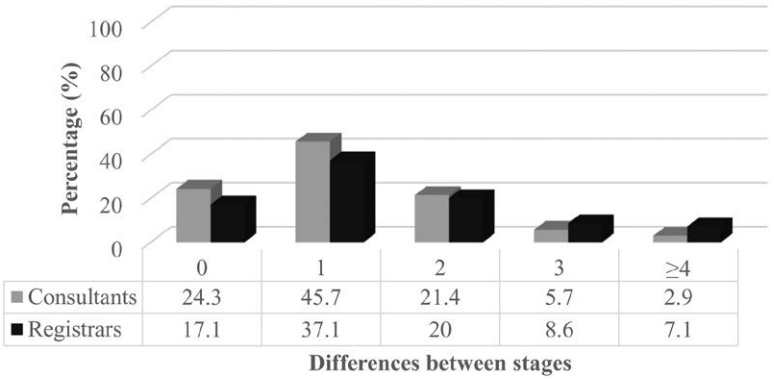
Copyright © 2020 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Fig. 2



All observers classified these images the same, (a) stage III and (b) stage VI.

Fig. 3



Percentage difference and staging distribution for consultants and registrars.

images were staged within a single stage of each other was stage II or III, where the difference in description is subjective in whether the medial defect has become a step or not. Of the four images staged, the same between the two surveys three of those were staged as stage III.

When separating out the residents from the consultants, there was a difference between the two groups with the consultants trending toward having a narrower distribution of results (Fig. 3), although this did not reach statistical significance (χ^2 = 8.942, P = 0.063). The κ-value looking at reliability ranged from 0.05 to 0.37 with the overall κ-value for all calculated reliabilities for all observers combined being 0.24 ($P < 0.01$). Separating out the consultants from the residents, the overall reliability κ for consultants was 0.25 ($P < 0.01$) and for the residents was 0.239 ($P < 0.01$)

Reproducibility

Survey 2 was used as a means of assessing the reproducibility by restaging the same images with the order changed. The observer with the most images staged the same was consultant 1, who staged 43 images the same, whereas the least images staged identically was by resident 1, who staged 22 images the same. Overall there was a mean of 34 images staged identically by the same

Copyright © 2020 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

observer between the two surveys with the consultants averaging 41.6 images staged identically opposed to the residents who averaged 27.6 images staged identically across the two surveys. The Cohens κ for the reproducibility ranged from 0.15 for resident 1 to 0.51 for consultant 1, with a mean overall reproducibility κ of 0.38. When separating the consultants and residents, the consultants had a mean κ-value of 0.48 whereas the residents had a mean κ of 0.26.

Discussion

Blount's disease is a rare condition affecting the proximal tibial physis and is divided into infantile, juvenile and adolescent subtypes [1]. Blount's disease is predominantly classified according to the Langenskiöld classification system, which was introduced in 1952 after a review of 23 cases of infantile tibia vara [2]. This classification system has been studied by numerous authors as a means of guiding treatment [4,11–13] with treatments ranging from observation and bracing to when the optimal time for surgery might be. Other authors have used the Langenskiöld classification as a means of prognostication [6–9]. Interobserver variability of the Langenskiöld classification has only been reviewed in one study [5], and that study found frequent discrepancies between observers and an overall poor agreement. The current study focused on both the reliability and reproducibility of the Langenskiöld classification.

There was an overall low percentage (2.9%) of images that were staged the same by all six observers in each survey, and 28.6% then 38.6% of the images the observers all differed by a single stage in the first and second surveys, respectively. Three of the four images stage the same by all observers were stage III, and where the stages differed by one the most common combination over both surveys was stage II and III. This suggests that stage III could possibly be the most reliable and recognizable stage with its description of the deep, step-like saucer shape. There was a trend to the consultants having less of a distribution in their results compared to the residents, although this was not statistically significant. This is comparable to what Stricker *et al.* [5] noted when they assessed the interobserver reliability of the Langenskiöld classification in a study of 60 knee radiographs, finding a higher correlation for their specialist staff compared to their residents. Differing from Stricker *et al.*'s study, however, where they noted the residents staging more images identically at 55% as opposed to their consultants 49%, we found that the consultants had a higher percentage of images staged identically with consultants staging 24.3% of the images the same compared to the residents 17.1%. Both studies did, however, find that there were more images within a single stage difference in the specialist group compared to the resident group, but as the difference reached more than three stages the residents had more images with this spread. The combined overall Fleiss Kappa was 0.24, which shows a fair agreement amongst observers. This is comparable to the findings of Stricker *et al.* [5]. In the current study, the Fleiss Kappa for consultants alone was slightly better at 0.25, whereas the Fleiss Kappa for residents was 0.24. All of these values fall into the category of fair agreement, although it can be appreciated that the consultants had a slightly better agreement than the residents.

To our knowledge this is the first study evaluating the reproducibility of the Langenskiöld classification. Of the 70 images staged over the two surveys, there was an average of 34 images that were staged identically amongst all six observers. The consultants were slightly more consistent with an average of 41.6 images staged the same while the residents had an average of 27.6 images staged the same. The κ statistic also mirrors the discrepancy between the two groups where the mean κ-value for the reproducibility overall showed a fair agreement at 0.38. Again, in separating the consultants from the residents showed the consultants had a higher mean κ at 0.48 to the residents 0.26. This showed the consultants had a moderate reproducibility while the residents only had fair reproducibility

In light of both the reliability and reproducibility showing only a fair correlation, below the desired 0.4, the practice at our institution relies more on the clinical presentation of the patient for decision making and placed very little emphasis on the Langenskiöld classification. Our choice of treatment relies more on the age of the patient and degree of deformity at presentation, with the majority undergoing a high tibial osteotomy. The degree of overcorrection is dependent on the age, rather than the Langenskiöld classification. Larger children with adolescent type Blount's disease may require a more gradual deformity correction using a circular ring fixator system. The decision regarding medial plateau elevation would be guided by the presence of a posteromedial clinical defect as detected by a lateral thrust when walking and a positive Siffert-Katz sign.

Limitations of this study were that some observers may have guessed the staging of some images without referring to the given reference description. This could affect the accuracy of the outcome, although it is accounted for in the κ calculation as opposed to the Pearson's correlation coefficient where it is not.

In conclusion, we found an overall fair level of reliability and reproducibility for the Langenskiold classification, with the κ for both falling below the generally acceptable threshold of 0.4, but with the consultants being more consistent than the residents. We recommend caution when using the Langenskiöld classification for any prognostication and decision making in the management of these children and that the patient be classified by the most senior clinician available to do so. We

Copyright © 2020 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

also recommend that caution be used when interpreting any studies relying on this classification for prognosis or treatment guidelines. Other classification systems such as the more objective medial slope angle or using other imaging modalities like MRI need to be developed and assessed for reliability and reproducibility, possibly paving the way for a new classification system to be used worldwide.

Acknowledgements

Conflicts of interest

There are no conflicts of interest

References

1 Birch JG. Blount disease. *J Am Acad Orthop Surg* 2013; **21**:408–418.
2 Langenskiold A. Tibia vara; (osteochondrosis deformans tibiae); a survey of 23 cases. *Acta Chir Scand* 1952; **103**:1–22.
3 Langenskiöld A. Tibia Vara. *J Pediatr Orthop* 1994; **14**:152–155.
4 LaMont LE, McIntosh AL, Jo CH, Birch JG, Johnston CE. Recurrence after surgical intervention for infantile tibia vara: assessment of a new modified classification. *J Pediatr Orthop* 2019; **39**:65–70.
5 Stricker SJ, Edwards PM, Tidwell MA. Langenskiöld classification of tibia vara: an assessment of interobserver variability. *J Pediatr Orthop* 1994; **14**:152–155.
6 Shinohara Y, Kamegaya M, Kuniyoshi K, Moriya H. Natural history of infantile tibia vara. *J Bone Joint Surg Br* 2002; **84**:263–268.
7 Laville JM, Chau E, Willemen L, Kohler R, Garin C. Blount's disease: classification and treatment. *J Pediatr Orthop B* 1999; **8**:19–25.
8 Khanfour AA. Does langenskiold staging have a good prognostic value in late onset tibia vara? *J Orthop Surg Res* 2012; **7**:23.
9 Hofmann A, Jones RE, Herring JA. Blount's disease after skeletal maturity. *J Bone Joint Surg Am* 1982; **64**:1004–1009.
10 Greene WB. Instructional course lectures, the American Academy of Orthopaedic Surgeons. Infantile tibia vara. *JBJS* 1993; **75**:130–143.
11 Richards BS, Katz DE, Sims JB. Effectiveness of brace treatment in early infantile Blount's disease. *J Pediatr Orthop* 1998; **18**:374–380.
12 Loder RT, Johnston CE 2nd. Infantile tibia vara. *J Pediatr Orthop* 1987; **7**:639–646.
13 Sabharwal S. Blount disease. *J Bone Joint Surg Am* 2009; **91**:1758–1776.
14 Langenskioeld A, Riska EB. Tibia vara (osteochondrosis deformans tibiae): a survey of seventy-one cases. *J Bone Joint Surg Am* 1964; **46**:1405–1420.

Copyright © 2020 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.