JOURNAL OF
CHILDREN'S ORTHOPAEDICS

Interobserver reliability of Diméglio and Pirani score and their subcomponents in the evaluation of idiopathic clubfoot in a clinical setting: a need for improved scoring systems

C. Bettuzzi
C. N. Abati
G. Salvatori
A. Zanardi
M. Lampasi

Abstract

Purpose Diméglio (DimS) and Pirani (PirS) scores are the most commonly used scoring systems for evaluation of clubfoot, with many centres performing both. Interobserver reliability of their global score has been rated high in a few studies, but agreement of their subcomponents has been poorly investigated. The aim of the study was to assess interrater reliability of global scores and of items in a clinical setting and to analyse overlapping features of the two scores.

Methods Fifty-six consecutive idiopathic clubfeet undergoing correction using the Ponseti method were independently evaluated at each casting session by two trained paediatric orthopaedic surgeons using both scores. Interobserver reliability of collected data was analysed; a kappa coefficient > 0.60 was considered adequate.

Results For DimS and PirS, the Pearson correlation coefficients were 0.87 and 0.91 (p < .0001) respectively, and kappa coefficients were 0.23 and 0.31. Among subcomponents, kappa values were rated > 0.60 only for equinus and curvature of lateral border in PirS; muscular abnormality in DimS was rated 0.74 but a high prevalence index (0.94) indicated influence of scarce prevalence of this feature. All other items showed k < 0.60 and were considered to be improved.

For overlapping features: posterior and medial crease showed similar agreement in the two systems, items describing equinus and midfoot adduction were much more reliable in PirS than in DimS.

Conclusions In a clinical setting, despite a high correlation of evaluations for total scores, the interobserver agreement of DimS and PirS was not adequate and only a few items were substantially reliable. Simultaneous use of two scores seemed redundant and some overlapping features showed different reliability according to criterion or scale used. Future scoring systems should improve these limitations.

Level of evidence: Level I – Diagnostic studies

Cite this article: Bettuzzi C, Abati CN, Salvatori G, Zanardi A, Lampasi M. Interobserver reliability of Diméglio and Pirani score and their subcomponents in the evaluation of idiopathic clubfoot in a clinical setting: a need for improved scoring systems. *J Child Orthop* 2019;13:478-485. DOI: 10.1302/1863-2548.13.190010

Keywords: clubfoot; Diméglio score; Pirani score; Ponseti method; reliability

Introduction

Scoring systems for evaluation of congenital clubfoot are meant to be a reliable tool, feasible and rapid to perform in clinical setting,[1,2] to be used as a guide to diagnosis, treatment and prognosis.[1,3] Commonly, they are used to distinguish between different severities of deformity, to monitor progression during correction and to help identify deformity relapse.[1,3] In addition, scores should ideally provide a prognostic contribution: they should anticipate the risk of relapse and relative time needed in foot abduction orthosis in the Ponseti method[4,5] and have a correlation with outcome. A large number of variables have been included in the systems proposed in the literature and it is still unclear which aspects are the most relevant.[1] The scoring systems that have been most commonly used in recent years for clinical and research purposes are those proposed by Diméglio[6] and Pirani,[7] constituted by the sum of specific clinical features of deformity and reducibility chosen by the authors. These scores, which are easy and rapid to perform, have been proven to fulfil some features of an ideal scoring system (ability to differentiate severity and to monitor correction and relapse) but poor evidence on other aspects (correlation with number of

Department of Paediatric Orthopaedics, Anna Meyer Children's Hospital, Florence, Italy

Correspondence should be sent to C. Bettuzzi, Anna Meyer Children's Hospital, Department of Paediatric Orthopaedics, Viale Pieraccini 24, 50139, Florence, Italy.
E-mail: camilla.bettuzzi@gmail.com

casts and need for Achilles tenotomy; long-term prognostic value) has been found,[5,8–10] which highlights the need to improve the systems used and understand which items of these scores could be retained in future systems.[1]

An essential feature of any evaluation is reliability. A score (or item) with low reliability compromises every further consideration or correlation. Both Diméglio (DimS) and Pirani (PirS) scores have been considered to have high interrater reliability in terms of total score,[2,5,7,10–17] but reliability of their subcomponents has been evaluated by very few studies.[5,13–17]

The purpose of the present study was to assess the reliability of PirS and DimS and of their items on the same sample of clubfeet and to suggest features and criteria that might be improved in future score systems.

Materials and methods

Patients with idiopathic clubfoot consecutively treated at our institution between November 2016 and August 2017 using the Ponseti method were prospectively enrolled. Parents provided informed consent and the local ethical committee approved the research. The study was performed in accordance with the ethical standards of the 1964 Declaration of Helsinki. Inclusion criteria were idiopathic clubfeet, age ≤ 4 months at first evaluation, no previous surgery. Exclusion criteria were age > 4 months, previous surgery, neurologic and syndromic clubfeet. All clubfeet were treated using the Ponseti method (including serial sessions of manipulation and casting, percutaneous Achilles tenotomy, if needed, and foot abduction orthosis).

Evaluations were performed using both PirS and DimS. PirS evaluates six clinical signs that characterize clubfoot.[7] Each of the six items is scored on a three-point scale (0 = none, 0.5 = moderate, 1 = severe abnormality). Three items compose the midfoot contracture score (MFCS): medial crease (MC-Pir), lateral part of head of talus (LHT) and curvature of lateral border (CLB). Another three compose the hindfoot contracture score (HFCS): posterior crease (PC-Pir), empty heel (EH) and rigid equinus (E-Pir). The MFCS and HFCS are then added together for total score ranging from 0 to 6, with a high score indicating a more severe deformity.

DimS is based on correction after applying a gentle reduction force.[6] Four parameters are evaluated: varus (VA), equinus (E-Dim), derotation of calcaneo-forefoot block (DER) and midfoot adduction (ADD). Each is scored on a five-point scale (0 to 4 points) resulting in a maximum score of 16 points for the most severe deformity. An additional four points result from the absence/presence (0/1 point) of four pejorative items: posterior crease (PC-Dim), medial crease (MC-Dim), cavus and abnormal musculature (MUSC). Once a total score (0 to 20 points) is calculated, feet are divided in four grades: grade I (0–5 points), II (6–10), III (11–15) and IV (16–20 points).

At each session (initial presentation and weekly treatment during casting) every foot underwent evaluation by two out of three paediatric orthopaedic surgeons (CB, CNA, ML) with experience in clubfoot treatment. No more than two evaluations per session were performed to avoid modifications in flexibility at last examination.[2,11,14,15] Evaluations were performed independently and in a blinded manner, since each examiner separately recorded the two scores and did not have information about the previous evaluations and findings of the other evaluator.

All evaluators had previous experience of at least six months with these scoring systems. Before the beginning of the study, a training session was held where all evaluators involved reviewed descriptions of both classifications; a score sheet with explanations, visual scales and pictures reporting degrees of reducibility was used to facilitate evaluations. For DimS, the original description of the score included marking the skin with a pencil and using a goniometer to support evaluation.[6] All the evaluators had previous experience with this way of applying the score in their learning curves (before beginning of the study) and then moved to an easier and more rapid application of the score with the support of visual scales and only occasional supplementation with a goniometer in case of doubt as reported by other authors[10,18,19] once they felt enough confidence with the score. This application was used for the study. Skin marking was deliberately not used for ethical reasons to avoid marking and erasing marks to allow blinded evaluations.

Data were analysed using Pearson correlation coefficient (PCC), p-value, percentage agreement (PA) and Cohen's kappa coefficient. PA was calculated by dividing observed agreement by total number of observations, which indicated how identical the repeated measurements were. Kappa coefficient is recommended to determine relative agreement between evaluators for nominal and categorical data, as it eliminates the effect of expected agreement at random. k ranges from ≤ 0.00, none, to 1.00, perfect agreement: values > 0.60 were interpreted as adequate, < 0.60 were considered inadequate and to be improved.

Interpretation of magnitude of kappa took into consideration the prevalence index that reflects prevalence of attributes in the sample. This was calculated by estimating the difference in proportion of agreement on positive and negative cases for the two raters, with values ranging from 0 (equal probability of positive and negative cases) to 1. If prevalence index is high (i.e. the prevalence of a positive rating is either very high or very low), chance agreement is also high and kappa should be reduced accordingly.[20]

Results

Thirty-five infants (31 boys, 4 girls) with 56 idiopathic clubfeet (21 bilateral, 14 unilateral; 29 right, 27 left) met criteria and were included in the study. Mean age at first evaluation in the study was 30 ± 25 days (range 8–125 days). A mean of 4.6 ± 1.2 casts (range 3–8) was performed. Achilles tenotomy was performed in 34 feet (97.1%) at a mean age of 60.3 ± 22.5 days (range 40–137). Complete correction was achieved in all cases. A total of 144 sessions (with two evaluations) were recorded. Due to organizational difficulties (not due to clinical problems), it was not possible for two observers to be present at every visit;[2] those sessions were not recorded. Patients' demographics and characteristics of evaluations are reported in Table 1.

Total scores

Statistical analysis (Table 2) found that PCC was 0.87 and 0.91 (p < .0001) for DimS (0 to 20 points) and PirS (0 to 6 points) respectively, whereas kappa coefficient was 0.23 and 0.31. Percentage agreement of total score for DimS was 29.9%, with score being within 1 point ('PA ± 1') in 71.5% of cases. PA of grades I to IV of DimS was 76.4%. For PirS, percentage agreement was 37.5%, with score being within 0.5 points ('PA ± 0.5') in 79.9% of cases.

Subparameters

In DimS (Table 3), k was > 0.60 only for MUSC (0.74). All other items were between 0.40 and 0.60. Percentage agreement was lower for parameters assessing reducibility

Table 1 Characteristics of the sample and of evaluations.

Gender	31 boys, 4 girls
Age at first cast (days)	26 ± 22 (range 8–118)
No. of casts performed	4.6 ± 1.2 (range 3–8)
% tenotomy	97.1% (34 feet)
Age at first evaluation (days)	30 ± 25 (range 8–125)
Assessments per foot	2.5 (range 1–5)
Last time point for assessment (days)*	17.6 (range 0–55)
Evaluations at first cast	38/144
Evaluations per rater	CB 109; CNA 75; ML 104

Note. CB, C. Bettuzzi; CNA, C.N. Abati; ML, M. Lampasi.
*time between last assessment and tenotomy or brace application

Table 2 Interobserver reliability of Diméglio and Pirani global scores in our study and revision of the literature.

Author, year	No. of sessions of evaluation (feet)	Professionals involved in the study (per session)	Score analysed	Statistical test used	Results
Current study	144 (56)	3 POSs (2 per session)	Diméglio	k PCC PA	0.23 0.87 (p < .0001) PA = 29.9%; PA ± 1 = 71.5%; PA for grades (I-IV) = 76.4%
			Pirani	k PCC PA	0.31 0.91 (p < .0001) PA = 37.5%; PA ± 0.5 = 79.9%
Lampasi et al, 2018	Not reported	POSs (not reported)	Diméglio	ICC	0.96 (p < .0005)
			Pirani	ICC	0.94 (p < .0005)
Sharma et al, 2018	115 (115)	1 OS, 1 resident doctor, 1 nonmedical counsellor (3 per session)	Pirani	Difference between the means	Difference between the means of total scores < 0.1
Jain et al, 2017	80 (80)	5 OSs (5 per session)	Pirani	k	0.71
Fan et al, 2017	250 (250)	1 POS, 1 radiologist (not clear in the manuscript)	Diméglio	ICC	0.81 for all feet 0.37–0.40 for feet with 2, 7, 8 casts 0.71–0.82 for feet with 3–6 casts
			Pirani	ICC	0.79 for all feet 0.37–0.40 for feet with 2, 7, 8 casts 0.73–0.88 for feet with 3–6 casts
Cosma and Vasilescu, 2015	Not reported (411)*	2 senior staff POSs	Diméglio	PCC PA	0.85 (p < .0001) PA = 39.17%; PA ± 1 = 64.23%
			Pirani	PCC PA	0.89 (p < .0001) PA = 23.84%; PA ± 0.5 = 70.06%
Harvey et al, 2014	65 (39)	19 PTs, experienced and novice (2 per session)	Pirani	ICC	0.90
Jillani et al, 2014	92 (92)	1 OS, 1 allied health worker	Pirani	k PA	0.362 PA = 41%
Shaheen et al, 2012	546 (91)	1 PT and 1 OS	Pirani	k	0.50
Pirani et al, 2008	Not reported	1 OS, 1 orthopaedic resident	Pirani	k	0.92
Wainwright et al, 2002	Not reported (13)	2 consultant POSs, 1 senior PT, 1 OS (2–4 per session)	Diméglio	k	0.77 for 2 consultants Moderate (0.41–0.60) for all observers
Flynn et al, 1998	55 (55)	2 POSs, 1 PT (2–3 per session)	Diméglio	PCC PA	0.83(p = 0.0001) PA = 20%; PA ± 1 = 61.8
			10-point Pirani**	PCC PA	0.90 (p = 0.0001) PA = 29.1%; PA ± 1 = 89%

Note. POS, paediatric orthopaedic surgeon; OS, orthopaedic surgeon; PT, physiotherapist; PCC, Pearson correlation coefficient; ICC, intraclass correlation coefficient; k, Cohen's kappa; PA, percentage agreement; PA ± 1, percentage agreement within 1 point between the two evaluators; PA ± 0.5, percentage agreement within 0.5 points between the two evaluators.
*age 7 days–13 years; **10-point Pirani = old version of PirS.

Table 3 Interobserver reliability of subcomponents of DimS in our study and revision of the literature.

ITEMS		ADD		VA		E-Dim		DER		MC-Dim			PC-Dim			Cavus			MUSC		
Studies		k	PA %	k	PA %	k	PA %	k	PA %	K	PA %	PI	k	PA %	PI	k	PA %	PI	k	PA %	PI
Current study	k	0.40	–	0.43	–	0.53	–	0.54	–	0.50	–	–	0.59	–	–	0.55	–	–	0.74	–	–
	PA	–	59.0	–	61.8	–	72.2	–	67.4	–	86.1	–	–	88.9	–	–	91.0	–	–	98.6	–
	PI	–	–	–	–	–	–	–	–	–	–	0.67	–	–	0.68	–	–	0.78	–	–	0.94
Lampasi et al, 2018	k	0.49	–	0.50	–	0.81	–	0.57	–	0.72	–	–	0.59	–	–	0.69	–	–	0.76	–	–

Note. ADD, midfoot adduction; VA, varus; E-Dim, equinus; DER, derotation of calcaneo-forefoot block; MC-Dim, medial crease; PC-Dim, posterior crease; MUSC, abnormal musculature; k, Cohen's kappa; PA, percentage agreement; PI, prevalence index.

Table 4 Interobserver reliability of subcomponents of PirS in our study and revision of the literature.

ITEMS		CLB		LHT		MC-Pir		PC-Pir		EH		E-Pir	
Studies		k	PA %	k	PA %	k	PA %	k	PA %	k	PA %	k	PA %
Current study	k	0.73	–	0.58	–	0.50	–	0.55	–	0.43	–	0.72	–
	PA %	–	83.3	–	72.9	–	77.8	–	74.3	–	70.8	–	86.1
Lampasi et al, 2018	k	0.84	–	0.58	–	0.69	–	0.65	–	0.50	–	0.58	–
Sharma et al, 2018	k	0.66	–	0.46	–	0.38	–	0.62	–	0.37	–	0.51	–
Jain et al, 2017	k	0.56	–	0.53	–	0.43	–	0.46	–	0.39	–	0.68	–
Harvey et al, 2014	ICC	0.82*	–	0.63*	–	0.76*	–	0.80*	–	0.57*	–	0.83*	–
Jillani et al, 2014	k	0.72	–	0.70	–	0.62	70	0.67	–	0.54	79.0	0.39	–
Shaheen et al, 2012	k	0.54	–	0.56	–	0.57	–	0.61	–	0.72	–	0.51	–
	PA %	–	86.8	–	80.2	–	74.7	–	90.1	–	93.4	–	78

Note. ICC, intraclass correlation coefficient; k, Cohen's kappa; PA, percentage agreement; CLB, curvature of lateral border; LHT, lateral part of head of talus; MC-Pir, medial crease; PC-Pir, posterior crease; EH, empty heel; E-Pir, rigid equinus.
*ICC

of deformity (with ADD and VA being the lowest at 59% and 61.8% respectively, and E-Dim the highest at 72.2%) than for pejorative items (ranging between 86.1% for MC-Dim and 98.6% for MUSC).

For a better understanding of the (apparently) high percentage agreement of these pejorative items and the high kappa of MUSC, we calculated number of feet rated as positive and the prevalence index. MUSC was rated positive by at least one observer only in five (3.5%) out of 144 sessions (and only in three there was concordance) and prevalence index was very high (0.94). For cavus, medial and posterior crease in DimS (22, 34 and 129 positive sessions, respectively) prevalence index was high (0.78, 0.67 and 0.68) as well. On the other hand, in PirS (Table 4) k was > 0.60 for E-Pir (0.72) and CLB (0.73). The other items had k values < 0.60, with empty heel (0.43) being the lowest.

Percentage agreement ranged from 70.8% for empty heel to 86.1% for E-Pir. For calculated scores, MFCS (52.1%) and HFCS (55.6%) showed very low percentage agreement. Empty heel was rated '1' by at least one observer in 115 out of 144 sessions (and there was concordance in only 74) and was rated '0' only in three sessions.

Discussion

Idiopathic clubfoot is mainly characterized by forefoot adduction, hindfoot varus, cavus and equinus[21] with a wide range of severity in terms of amount and reducibility of the deformity. Many efforts have been made to improve understanding of the behaviour of clubfeet with different severity and features.[3,22] To this end, clubfeet have been distinguished by means of classifications or by evaluation systems.[1] With the first tool, feet are graded considering basic features that do not change with time (mainly aetiology). Conversely, evaluation systems describe aspects (deformity, shape, range of motion, rigidity or radiographic angles) that are evident at the time of presentation but change with time and treatment.[1,17] Both are important to understand deformity, guide treatment, compare results and, hopefully, to get prognostic indications.

At our institution, we are performing a prospective study with the aim of improving current classification and evaluation systems. Different systems have been proposed in the past and many have been abandoned or have not found diffusion due to intrinsic limitations (complexity, low reliability, need for radiographs, too many descriptive and qualitative features, etc).[11] At present, DimS[6] and PirS[7] have overcome most of the limitations of previous systems and have reached a worldwide diffusion. Many centres are using both to allow comparison of results with other centres, since the superiority of one or the other has not been proven.

Reliability of global scores (PirS and DimS)

Reliability of the two scores has been evaluated in few previous studies[2,5,7,10–17] (Table 2) and considered high in most of them. It should be stressed[12,15] that some studies are not independent from the ideators of the scores and that results are not perfectly comparable since different professionals and levels of experience were involved and different methods were used (inclusion of clubfeet throughout treatment or not, inclusion of older children, use of goniometer, etc). In addition, different statistical tests were used: PCC and intraclass correlation coefficients > 0.80 have been reported in several papers,[5,10–12,15] whereas low values of kappa coefficient and percentage agreement have been reported as well.[11,12,16,17] Fan et al[10] reported that despite a high reliability for initial DimS and PirS in the whole sample, reliability was poor or moderate for feet with low or high number of casts. Also, the interpretation of results was sometimes questionable, for example with PirS considered reliable despite k values of 0.362[16] or 0.50[17] or PA ± 0.5 = 70.06%[12] (which means about 30% of feet being rated with at least 1 PirS point of difference).

In our work, we have used different statistical tests to allow for comparison with previous studies and our results were in line with most of them showing that PCC was high (0.91 for PirS and 0.87 for DimS); yet, PCC is best used for correlation and not for agreement. As for agreement, k coefficient and percentage agreement were found to be inadequate: our findings showed, for example, that feet were rated with at least 1 PirS point of difference in one case out of every five, and with a different grade of DimS in one case out of every four, which is remarkable, considering that some authors[4,5,23] suggest using these scores for therapeutic decisions such as tenotomy or time needed in abduction brace.

Reliability of subparameters

Reliability of items included in the scores has been studied in very few previous papers (Tables 3 and 4) with the same limitations described for global scores. Many comments have been based on criteria for interpretations of kappa values proposed by Landis and Koch,[24] that led previous authors to judge items as adequate even if rated < 0.60.[17] Yet, interpretation of kappa values is not straightforward and, for health research, values < 0.60 may indicate inadequate agreement.[25] Our study showed adequate k (> 0.60) only for equinus and curvature of lateral border in PirS and abnormal musculature (MUSC) in DimS: all the others were considered to be improved.

Interpretation of a given magnitude of kappa is influenced by other factors that have not been analysed (or just mentioned)[15] in previous papers. Analysis of prevalence index clearly showed that high percentage agreement of pejorative items in DimS and the high kappa of MUSC are in fact secondary to very asymmetrical distribution of features in the sample. In fact, the item describing abnormal musculature (MUSC) is vague[9,11] and includes too many different entities (muscle fibrosis, aplasia, hypertonia, imbalance, stiffness, fat infiltration, etc)[9] that increase the risk of disagreement. The same considerations can be given to other uncommon features like cavus that are not easy to rate despite high percentage agreement (prevalence index 0.78).

On the other hand, other features (posterior crease in DimS = 1, empty heel = 1) are very common, because these are very frequent characteristics of clubfoot at presentation[9,23] and in most cases persist until tenotomy with a sudden correction thereafter.[9,23] Since studies on the reliability of clubfoot scores typically include more feet in the corrective phase than feet at end of treatment, an apparent increase of reliability may result.

In contrast, this construct of study (including clubfeet throughout correction) may highlight the limitations of items that have a progressive evolution of correction in Ponseti method (varus, adduction, derotation, equinus in DimS, LHT in PirS).[9,23] In fact, in our study these parameters showed low reliability values and this could be due to scarce definition of their points of transition: for example, limits between points of varus, adduction, derotation and equinus in DimS are easily misleading with different force exerted. Also LHT seemed to be a quite complex feature to assess, with scarce definition of points of transition.[12]

A low reliability was found also for empty heel, as previously reported.[11,12] Some strategies have been suggested[15] to improve reliability, such as comparing palpation of the heel with the feel of touching a chin, nose or cheek to reflect normal, moderate and severe scores respectively. We have used these modifications in our clinical practice, but reliability results were still low. In our opinion this parameter has very limited clinical utility: in our practice, we focus on emptiness of the heel only at the time we decide whether to perform tenotomy or not, if all other components are corrected, a dorsiflexion ≥ 10–15° is achieved and we suspect an incomplete correction of equinus of the calcaneus due to persistence of empty heel or posterior crease; in this case, we perform a lateral radiograph of the foot in maximal dorsiflexion, as previously suggested.[26]

Items that overlap between PirS and DimS

In general, PirS and DimS are different in nature, since DimS is based on *degrees* of 'reducibility without forcing the foot'[6] and PirS considers the *morphologic aspect*[12] or specific *physical findings* in a 'gently corrected position' or with the foot corrected 'as much as possible'.[7] Yet, the two scores share evaluation of some clinical features that are exactly the same (posterior crease, medial crease, equinus) even though rated differently (two- versus three-point scale or three- versus five-point scale) or that are complementary (for example, CLB and ADD are both evaluating adduction in a different manner),

In our routine experience we regularly perform both scores as recommended by some authors,[12] but feel that this simultaneous use is, at best, redundant due to these overlapping items. Performing a direct comparison of these items has some limitations, given the intrinsic difficulty in interpretation of kappa, the different criteria of the two scores and different subdivision of items, but some considerations can be made.

For items describing posterior and medial creases, availability (in PirS) of a three-point scale apparently facilitates description of intermediate conditions in comparison with the two-point scale of DimS, but this did not lead to an increase of k, at least, not one that was adequate: posterior crease (0.55 and 0.59 respectively) and medial crease (0.50 in both) showed similar agreement in the two scores.

For items describing adduction of clubfoot, a great difference was evident between ADD in DimS (0.40) and CLB in PirS (0.73). The criterion (morphologic aspect) and the scale used (three-point) in PirS led to higher agreement: low agreement in DimS is likely due to different stretching forces applied by evaluators on a five-point scale.

Similarly, for items describing equinus, reliability of equinus in PirS (0.72) was rated much higher than equinus in DimS (0.53) and this was likely due to a more frequent disagreement between raters at points of transition between two grades if a five-point scale is used for equinus.

Considerations for future scoring systems

Efforts should be made to create new scoring systems to combine advantages of the two scores, maintaining those items whose reliability and clinical utility have been proven,[5] and eliminating useless overlapping. Some items have scarce therapeutic (or prognostic) value or inadequate agreement (for example, empty heel) and should be reconsidered. Other items have clinical utility (for example, adduction, derotation or equinus) and require strategies to improve their reliability (changing criteria or scales or improving definition of points of transition): for example, Harvey et al[15] tried to modify the three-point scale for PirS items into a five-point scale by introducing an expanded certainty measure and this achieved increased reliability. Features such as posterior crease may seem to have an adequate reliability[5,15] due to their frequent representation in the samples:[9] the decision to maintain or eliminate these items should be based on primarily their clinical ability to differentiate severity or prognosis, which does not seem to be evident. Features such as muscular abnormalities (that have been related to resistance to treatment and to different behaviour during treatment)[9] are clearly to be retained, but will hardly show statistical correlations unless reliability is improved.

Strengths and limitations

The strengths of this study are that it represents an independent work, with prospective and consecutive data collection and inclusion of only paediatric orthopaedic surgeons (poorly investigated in previous studies),[14] who in many centres are responsible for evaluations. Exclusive involvement of clinicians with experience in evaluation and treatment of clubfoot is at the same time a strength (uniformity of evaluators) and a limitation: inclusion of less experienced healthcare professionals could have provided more comprehensive results. Similarly, concurrent evaluation of feet with two scores by the same operator provides a strength (results of both scores are influenced by operator- and sample-dependent variables in the same way: this allows direct comparison of the two scores, particularly for overlapping items, which has never been reported, to our knowledge) and a bias, as preliminary evaluation with one score could have influenced the second one.

A possible limitation was the methodological application of DimS we used, that did not perfectly correspond to the original description of the score.[6] The simplified application we used (by visual estimation with support of visual scales and only occasional supplementation by goniometer in case of doubt) has been used by other previous independent papers as well[10,12,18,19] (even in papers analysing the reliability of DimS[10]). A complete and strict application (marking the skin and using goniometers for all evaluations)[6] is very useful in the learning curve phase but is not regularly applied by many centres in daily practice. Besides, use of goniometer in other papers[11] led to findings of reliability for DimS very similar to ours. We acknowledge that this is a limitation, but our aim was to replicate the setting that is more commonly applied in clinical practice.

Another limitation is that we included in the study evaluations performed throughout the corrective casting sessions, whereas the scores were originally intended for initial assessment.[6,7] Yet, with time the applications of the scores have been extended[19,23] and other authors have also examined reliability using similar methods.[2,12,15,17]

As regards the evaluation of individual subparameters, it could be argued that the scores were designed to be used as a global assessment of the foot and should not be divided, but this refers to the clinical value of the score, whereas adequate reliability of subparameters is required anyhow.

In conclusion, in a clinical setting the interobserver agreement of DimS and PirS was not adequate and only a few items were substantially reliable. Simultaneous use of two scores seemed redundant and some features evaluated by both scores showed different reliability according to the criterion or scale used. Further studies are required to evaluate the prognostic value of the items.

Received 29 January 2019; accepted after revision 10 September 2019.

COMPLIANCE WITH ETHICAL STANDARDS

FUNDING STATEMENT

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

OA LICENCE TEXT

ETHICAL STATEMENT

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.
Informed consent: Not required for this work.

ICMJE CONFLICT OF INTEREST STATEMENT

None declared.

ACKNOWLEDGEMENTS

The authors wish to thank Prof Giancarlo Bettuzzi and Dr Salvatore De Masi for help with statistical analysis.

AUTHOR CONTRIBUTIONS

CB: Data acquisition, analysis and interpretation of data, ideation and writing of the manuscript.
CNA: Data acquisition, critical revision of the manuscript.
GS: Data acquisition, critical revision of the manuscript.
AZ: Data acquisition, critical revision of the manuscript.
ML: Data acquisition, analysis and interpretation of data, ideation and writing of the manuscript.

REFERENCES

1. **Chu A, Labar AS, Sala DA, van Bosse HJ, Lehman WB.** Clubfoot classification: correlation with Ponseti cast treatment. *J Pediatr Orthop* 2010;30:695–699.

2. **Wainwright AM, Auld T, Benson MK, Theologis TN.** The classification of congenital talipes equinovarus. *J Bone Joint Surg Br* 2002;84:1020–1024.

3. **Goriainov V, Judd J, Uglow M.** Does the Pirani score predict relapse in clubfoot? *J Child Orthop* 2010;4:439–444.

4. **Zhao D, Liu J, Zhao L, Wu Z.** Relapse of clubfoot after treatment with the Ponseti method and the function of the foot abduction orthosis. *Clin Orthop Surg* 2014;6:245–252.

5. **Lampasi M, Abati CN, Bettuzzi C, Stilli S, Trisolino G.** Comparison of Dimeglio and Pirani score in predicting number of casts and need for tenotomy in clubfoot correction using the Ponseti method. *Int Orthop* 2018;42:2429–2436.

6. **Diméglio A, Bensahel H, Souchet P, Mazeau P, Bonnet F.** Classification of clubfoot. *J Pediatr Orthop B* 1995;4:129–136.

7. Pirani S, Hodges D, Sekeramyi F. A reliable and valid method of assessing the amount of deformity in the congenital clubfoot deformity. *J Bone Joint Surg Br* 2008;90(suppl):53.

8. **Dobbs MB, Rudzki JR, Purcell DB, Walton T, Porter KR, Gurnett CA.** Factors predictive of outcome after use of the Ponseti method for the treatment of idiopathic clubfeet. *J Bone Joint Surg Am* 2004;86:22–27.

9. **Lampasi M, Trisolino G, Abati CN, et al.** Evolution of clubfoot deformity and muscle abnormality in the Ponseti method: evaluation with the Dimeglio score. *Int Orthop* 2016;40:2199–2205.

10. **Fan H, Liu Y, Zhao L, et al.** The correlation of Pirani and Dimeglio scoring systems for Ponseti management at different levels of deformity severity. *Sci Rep* 2017;7:14578.

11. **Flynn JM, Donohoe M, Mackenzie WG.** An independent assessment of two clubfoot-classification systems. *J Pediatr Orthop* 1998;18:323–327.

12. **Cosma D, Vasilescu DE.** A clinical evaluation of the Pirani and Dimeglio idiopathic clubfoot classifications. *J Foot Ankle Surg* 2015;54:582–585.

13. **Sharma P, Verma R, Gaur S.** Interobserver reliability of Pirani clubfoot severity scoring between an orthopedic surgeon, a resident doctor, and a nonmedical counsellor at a clubfoot clinic. *Indian J Orthop* 2018;52:645–650.

14. **Jain S, Ajmera A, Solanki M, Verma A.** Interobserver variability in Pirani clubfoot severity scoring system between the orthopedic surgeons. *Indian J Orthop* 2017;51:81–85.

15. **Harvey NJ, Mudge AJ, Daley DT, Sims SK, Adams RD.** Interrater reliability of physiotherapists using the Pirani scoring system for clubfoot: comparison with a modified five-point scale. *J Pediatr Orthop B* 2014;23:493–500.

16. **Jillani SA, Aslam MZ, Chinoy MA, Khan MA, Saleem A, Ahmed SK.** A comparison between orthopedic surgeon and allied health worker in pirani score. *J Pak Med Assoc* 2014;64:S127–S130.

17. **Shaheen S, Jaiballa H, Pirani S.** Interobserver reliability in Pirani clubfoot severity scoring between a paediatric orthopaedic surgeon and a physiotherapy assistant. *J Pediatr Orthop B* 2012;21:366–368.

18. **Marchal C, André-Vert J.** Fiche d'évaluation du pied bot varus équin congénital selon la classification de Diméglio. *Kinesither Rev* 2006;6:35–36. Article in French.

19. **Chaudhry S, Chu A, Labar AS, Sala DA, van Bosse HJ, Lehman WB.** Progression of idiopathic clubfoot correction using the Ponseti method. *J Pediatr Orthop B* 2012;21:73–78.

20. **Sim J, Wright CC.** The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257–268.

21. **Ponseti IV.** *Congenital clubfoot: fundamentals of treatment*. Oxford: Oxford University Press, 1996.

22. **Agarwal A, Gupta N.** Does initial Pirani score and age influence number of Ponseti casts in children? *Int Orthop* 2014;38:569–572.

23. **Lampasi M, Abati CN, Stilli S, Trisolino G.** Use of the Pirani score in monitoring progression of correction and in guiding indications for tenotomy in the Ponseti method: are we coming to the same decisions? *J Orthop Surg (Hong Kong)* 2017;25:2309499017713916.

24. **Landis GR, Koch GG.** The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174

25. **McHugh ML.** Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22:276–282.

26. **Kang S, Park SS.** Lateral tibiocalcaneal angle as a determinant for percutaneous Achilles tenotomy for idiopathic clubfeet. *J Bone Joint Surg Am* 2015;97:1246–1254.