Eur Radiol (2016) 26:1963–1970
DOI 10.1007/s00330-015-4013-8

MUSCULOSKELETAL

Scoring haemophilic arthropathy on X-rays: improving inter- and intra-observer reliability and agreement using a consensus atlas

Wouter Foppen[1] · Irene C. van der Schaaf[1] · Frederik J. A. Beek[1] · Helena M. Verkooijen[1,2] · Kathelijn Fischer[2,3]

Received: 20 May 2015 / Revised: 5 August 2015 / Accepted: 4 September 2015 / Published online: 24 September 2015

Abstract

Objectives The radiological Pettersson score (PS) is widely applied for classification of arthropathy to evaluate costly haemophilia treatment. This study aims to assess and improve inter- and intra-observer reliability and agreement of the PS.

Methods Two series of X-rays (bilateral elbows, knees, and ankles) of 10 haemophilia patients (120 joints) with haemophilic arthropathy were scored by three observers according to the PS (maximum score 13/joint). Subsequently, (dis-)agreement in scoring was discussed until consensus. Example images were collected in an atlas. Thereafter, second series of 120 joints were scored using the atlas. One observer rescored the second series after three months. Reliability was assessed by intraclass correlation coefficients (ICC), agreement by limits of agreement (LoA).

Results Median Pettersson score at joint level (PS_{joint}) of affected joints was 6 (interquartile range 3–9). Using the consensus atlas, inter-observer reliability of the PS_{joint} improved significantly from 0.94 (95 % confidence interval (CI) 0.91–0.96) to 0.97 (CI 0.96–0.98). LoA improved from ±1.7 to ±1.1 for the PS_{joint}. Therefore, true differences in arthropathy were differences in the PS_{joint} of >2 points. Intra-observer reliability of the PS_{joint} was 0.98 (CI 0.97–0.98), intra-observer LoA were ±0.9 points.

Conclusions Reliability and agreement of the PS improved by using a consensus atlas.

Key Points

- *Reliability of the Pettersson score significantly improved using the consensus atlas.*
- *The presented consensus atlas improved the agreement among observers.*
- *The consensus atlas could be recommended to obtain a reproducible Pettersson score.*

Keywords Haemophilia · Arthropathy · X-rays · Reliability · Agreement

Abbreviations

CI	95 % confidence interval
GRASS	Guidelines for reporting reliability and agreement studies
ICC	Intraclass correlation coefficient
IQR	Inter quartile range
LoA	Limits of agreement
MRI	Magnetic resonance imaging
PS	Pettersson score
PS_{joint}	Pettersson score at joint level
$PS_{patient}$	Pettersson score at patient level

Introduction

Severe haemophilia is characterized by spontaneous or trauma-related joint bleeds. Recurrent joint bleeds eventually result in progressive arthropathy through metabolic and

Electronic supplementary material The online version of this article (doi:10.1007/s00330-015-4013-8) contains supplementary material, which is available to authorized users.

✉ Wouter Foppen
w.foppen@umcutrecht.nl

[1] Department of Radiology, University Medical Center Utrecht, HP E01.132, Post Office Box 85500, 3508 GA Utrecht, The Netherlands

[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[3] Van Creveldkliniek, Department of Hematology, University Medical Center Utrecht, Utrecht, The Netherlands

mechanical joint destruction [1]. The radiological Pettersson score (PS) was already designed in 1980 and is widely applied to classify the osteochondral changes of haemophilic arthropathy in elbows, knees, and ankles [2]. It is used to evaluate the effects of different treatment strategies, especially in international studies [2–4]. The PS is an additive score consisting of eight items (Table 1). Some of these items leave room for subjective interpretation. This is not a problem if all joints are scored by a single observer. However, a reproducible PS is especially important in comparative international studies involving multiple observers focusing on small differences. These comparative international studies are performed to justify the high costs of clotting factor replacement therapy. Although magnetic resonance imaging (MRI) is the most sensitive imaging modality for evaluation of arthropathy, it is too costly and too time consuming for use in routine follow up. In contrast, routine plain X-rays of the main joints at 5-year intervals are recommended in routine follow-up of patients with severe haemophilia [5]. As conventional radiography is cheap and universally available, it is a useful tool for international comparisons of different clotting factor replacement strategies.

The reproducibility of a scoring method involves the reliability and agreement. Reliability is a measure to define how well patients can be differentiated with the tool of interest. Agreement is the extent in which scores by different observers are identical [6, 7]. Although the PS is available for over 30 years, only three studies have assessed its reproducibility [8–10]. Formal assessment of the limits of agreement, however, was not performed. Yet, these limits of agreement are important for the clinical interpretation of the results. Without knowledge on the reproducibility of the PS, it is not clear whether small differences in scores represent real differences in arthropathy, or are attributable to inter-observer variation.

The original paper about the PS describes the genesis of the score, including the radiological changes, which are included in the score. However, definitions or examples of the included radiological items were not provided. Improved definition of PS items is expected to improve reliability and agreement [8, 9]. An atlas with reference images of different stages of haemophilic arthropathy of different joints and corresponding PS could potentially be a helpful tool. The purpose of this research was to develop a consensus atlas for the PS and to evaluate the impact of this atlas on the inter-observer and intra-observer reliability and agreement.

Table 1 Classification of haemophilic arthropathy according to the Pettersson score [2]

Osteoporosis	Absent	0
	Present	1
Enlargement of epiphysis	Absent	0
	Present	1
Irregularity of subchondral surface	Absent	0
	Partially involved	1
	Totally involved	2
Narrowing of joint space	Absent	0
	Joint space>1 mm	1
	Joint space<1 mm	2
Subchondral cysts formation	Absent	0
	1 cyst	1
	>1 cyst	2
Erosion of joint margins	Absent	0
	Present	1
Gross incongruence of articulating bone ends	Absent	0
	Slight	1
	Pronounced	2
Joint deformity (angulation and/or displacement)	Absent	0
	Slight	1
	Pronounced	2
Total joint score (max 13 points)		

Materials and methods

This study was conducted according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [7]. Inter-observer and intra-observer reproducibility of the Pettersson score at joint level (PS_{joint}) and the Pettersson score at patient level ($PS_{patient}$) were assessed in this study. The study was approved by the institutional ethical review board, and informed consent was waived.

Pettersson score

The PS is based on typical findings of haemophilic arthropathy on posterior-anterior and lateral X-rays, including osteoporosis, enlargement of epiphysis, irregularity of subchondral surface, narrowing of joint space, subchondral cysts, erosions at joint margins, incongruence between joint surfaces, and the angulation and/or displacement of articulating bone ends (Table 1). The maximum PS_{joint} is 13 points. The $PS_{patient}$ represents the sum of the six joints (elbows, knees, and ankles) with a maximum score of 78 points.

Sample size

Although determining sample size for studies on reliability and agreement is not straightforward [7], Shoukri and colleagues provided approximate sample sizes depending on the expected reliability values and the number of observers. For two observers and expected intraclass correlation coefficients larger than 0.80, inclusion of about 50 joints are needed [11], as about 50 % of joints of haemophilia patients are

affected on X-rays [12]. A sample size of ≥ 100 joints in the present study would suffice.

Patients and observers

Two series (mean time between series: 5.6 years) of X-rays (bilateral elbows, knees, and ankles) of 20 patients representing the full range of radiological variation were scored in this study. These patients were included based on previous PS by an experienced observer [13]. The current study involved three observers with different levels of experience in using the PS: one radiologist with experience using the PS (scoring>450 joints previously [3]), one radiologist without experience using the PS, and a medical doctor (PhD candidate on imaging of haemophilic arthropathy) without experience using the PS.

The reproducibility prior to the consensus atlas was evaluated by assessing the first series of X-rays (ten patients; 120 joints). These X-rays were scored independently by three observers. Ankylosis, arthroplasty, or arthrodesis were scored as a PS at joint level of 13. Agreement and disagreement in PS of the first series of X-rays were discussed in three consensus meetings. Common disagreements were discussed until consensus was reached. Example images with descriptions of the PS items were collected into a consensus atlas. Subsequently, the effect of the consensus atlas was evaluated: the same three observers independently scored the second series of X-rays (ten different patients; 120 different joints) with use of the developed consensus atlas.

To assess the intra-observer reliability and agreement while using the consensus atlas, the second series of X-rays were rescored after 3 months by the medical doctor without experience using the PS, blinded for the first results.

Analysis

The Mann-Whitney U test and Chi-squared test (two-sided) were used to compare distributions of patient characteristics between the first series of X-rays scored prior to the consensus atlas and the second series of X-rays scored with the consensus atlas. Generalized estimating equations (GEE) were performed to test whether PS were statistically different between the series of X-rays and to correct for repeated measurements and clustering of joints within patients. Due to variation of data distributions, three models were used: a binary model for comparison of percentages of abnormal scores, a gamma with log link model for comparison of medians across all scores, and a linear model for comparison of medians of abnormal scores only. The *p* values represent the adjusted analyses of the difference between the two series of X-rays. *P* values less than 0.05 were considered statistically significant.

Inter-observer and intra-observer reliability were assessed by two-way random intraclass correlation coefficient$_{agreement}$ (ICC) with 95 % confidence intervals (CI) for both the PS_{joint} and the $PS_{patient}$. The ICC is a measure to evaluate the relation between the variance in subjects and the variance in scores caused by the different observers. An ICC of 0.00 means a poor reliability among observers; an ICC of 1.00 means a perfect reliability. The reliability of the PS was considered significantly different in case the CI regarding the PS prior to the consensus atlas was not overlapping the ICC of the PS with use of the consensus atlas.

Agreement was assessed by a graphical method for multiple observers in a single plot according to Jones and colleagues [14], which is comparable to a Bland-Altman plot for two observers [14, 15]. The difference between each observer and the overall mean of all observers was calculated including the limits of agreement from the mean (LoA) (±2 standard deviations). The difference from the mean for each observer was subsequently plotted. The minimal detectable difference in PS can be interpreted as the change in PS beyond the maximal variance in scores caused by different observers. For the used method to assess agreement, the range between the higher and lower LoA (twice the LoA) could be interpreted as the minimal detectable difference. Analyses were performed in SPSS (IBM SPSS Statistics version 20, Armank, NY).

All six imaged joints of patients were included the analyses and results regarding reliability and agreement. Though most patients with haemophilia only have only several joint with frequent bleeds and subsequent arthropathy, about 50 % of joints were affected [12]. Healthy joints are easy to score. For that reason, a subgroup analysis of abnormal PS was performed to study the reliability and agreement of these affected joints only. An abnormal PS was defined as a PS>0 according to the mean of the three observers.

Results

X-rays of 240 joints in total from 20 patients were assessed. Joints with the whole range of haemophilic arthropathy were scored (PS_{joint} 0–13 points). Median age of patients was 35.4 years (inter quartile ranges (IQR) 28.6–46.1). Based on serum coagulation factor VIII / IX levels, fourteen patients had severe haemophilia A (FVIII <1 IU dL^{-1}), three had moderate haemophilia A (FVIII 2-5 IU dL^{-1}), two had severe haemophilia B (FIX <1 IU dL^{-1}), and one patient had moderate haemophilia B (FIX 2-5 IU dL^{-1}). Scored joints included three ankles with previous joint surgery (arthrodesis) (Table 2). A consensus atlas with example images and descriptions of the items (e.g., Fig. 1) was established and is now available as Electronic Supplementary Material.

Table 2 Patient and joint characteristics of the series of X-rays scored without the consensus atlas and the series of X-rays scored with the consensus atlas

	Series without consensus atlas (10 patients, 120 joints)	Series with consensus atlas (10 patients, 120 joints)	*p* value
Age (years)	35.8 (27.7–47.5)	33.7 (28.8–43.4)	0.97
Type of haemophilia (A / B)	90 % / 10 %	80 % / 20 %	1.00
Severity (moderate / severe) [a]	20 % / 80 %	20 % / 80 %	1.00
Treatment (on demand / prophylactic)	40 % / 60 %	10 % / 90 %	0.30
Joints with surgery	1.7 %	3.3 %	1.00
Pettersson score at joint level			
Percentage abnormal scores	64 %	49 %	0.28 [b]
Median of all scores	3 (0–7)	0 (0–4)	0.45 [b]
Median of abnormal scores only	6 (3–9)	4 (2–9)	0.67 [b]
Pettersson score at patient level			
Percentage abnormal scores	90 %	85 %	0.73 [b]
Median of all scores	25 (10–36)	9 (2–23)	0.28 [b]
Median of abnormal scores only	26 (19–37)	12 (7–24)	0.24 [b]

Values are percentages and medians (inter quartile ranges) regarding the average score of the three observers

[a] Severity of disease according to true serum levels of coagulation factors VIII in haemophilia A or IX in haemophilia B

[b] Calculated using regression analyses adjusted for repeated measurements and clustering of joints within patients

Overall, median PS_{joint} was 1 (IQR 0–7) and the median $PS_{patient}$ was 19 (IQR 7–28). A PS_{joint} above 0 was observed in 136/240 (57 %) joints and a $PS_{patient}$ above 0 in 35/40 (88 %) patients. The first series of X-rays (scored prior to the consensus atlas) included more joints with a positive PS (64 % with $PS_{joint}>0$) compared to the second series of X-rays, which was scored with the consensus atlas (49 % with $PS_{joint}>0$). A subgroup analysis of the abnormal Pettersson scores only is provided at the end of the results section. Patient characteristics and the range of scores were comparable among both series of X-rays.

Reliability

Inter-observer reliability (reflected by the ICC) of the PS_{joint} significantly improved from 0.94 (CI 0.91–0.96) to 0.97 (CI 0.96–0.98) with use of the consensus atlas (Table 3). For the $PS_{patient}$, the reliability significantly improved from 0.94 (CI 0.86–0.98) to 0.99 (CI 0.97–1.00).

The intra-observer reliability, which was only assessed with use of the consensus atlas, was excellent; 0.98 (CI 0.97–0.98) for the PS_{joint} and 0.99 (CI 0.98–1.00) for the $PS_{patient}$.

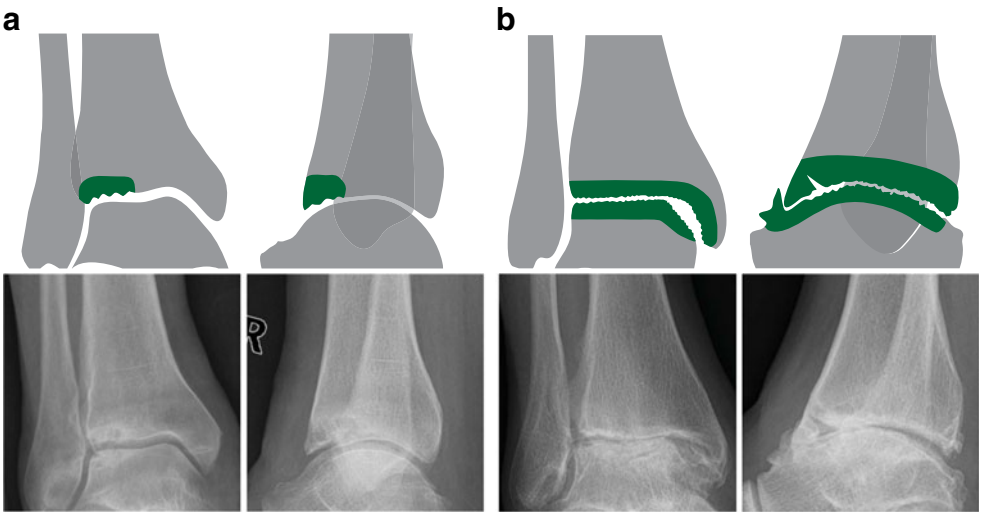

Fig. 1 Example images from the consensus atlas of 'irregularity of subchondral surface' in the ankle. *A*) Partially involved, minor subchondral cysts present as well. *B*) Totally involved, other significant joint alterations present as well

Table 3 Inter-observer and intra-observer reliability without the consensus atlas, and with use of the consensus atlas regarding the Pettersson score at joint level and the Pettersson score at patient level

	Pettersson score at joint level		Pettersson score at patient level	
	Without atlas	With atlas	Without atlas	With atlas
Results regarding all Pettersson scores				
Inter-observer	0.94 (0.91–0.96)	0.97 (0.96–0.98)*	0.94 (0.86–0.98)	0.99 (0.97–1.00)*
Intra-observer	–	0.98 (0.97–0.98)	–	0.99 (0.98–1.00)
Subgroup analysis of abnormal Pettersson scores only				
Inter-observer	0.88 (0.81–0.92)	0.94 (0.91–0.96)*	0.92 (0.80–0.97)	0.99 (0.97–0.99)*
Intra-observer	–	0.96 (0.93–0.97)	–	0.99 (0.97–1.00)

Values represent the intraclass correlation coefficients (95 % confidence intervals).

* Significant improvement of the reliability with use of the consensus atlas

Agreement

Before establishment of the consensus atlas, the experienced radiologist deviated the least from the mean score of the three observers (Figs. 2 and 3). Agreement among the observers as expressed by inter-observer LoA before the consensus atlas was ±1.7 points for the PS_{joint} (maximum score 13). The observers deviated less from the mean score after using the consensus atlas: the LoA for the PS_{joint} improved to ±1.1 points with use of the consensus atlas (Table 4). Regarding the intra-observer agreement after the consensus atlas, the LoA were ±0.9 points for PS_{joint}. Therefore, the minimal detectable difference for the PS_{joint} could be interpreted as > 2 points (rounded) after the consensus atlas. Thus, differences in PS_{joint}>2 points are attributable to true differences in arthropathy in case the PS_{joint} is scored with use of the consensus atlas.

For the $PS_{patient}$ (maximum score 78), the inter-observer LoA improved from ±6.5 points prior to the consensus atlas to ±3.2 points with use of the consensus atlas (Table 4). The intra-observer LoA after the consensus atlas were ± 2.4 points. In other words, the minimal detectable difference of the $PS_{patient}$ scored by different observers halved with use of the consensus atlas from > 13 points (rounded) to > 6 points (rounded).

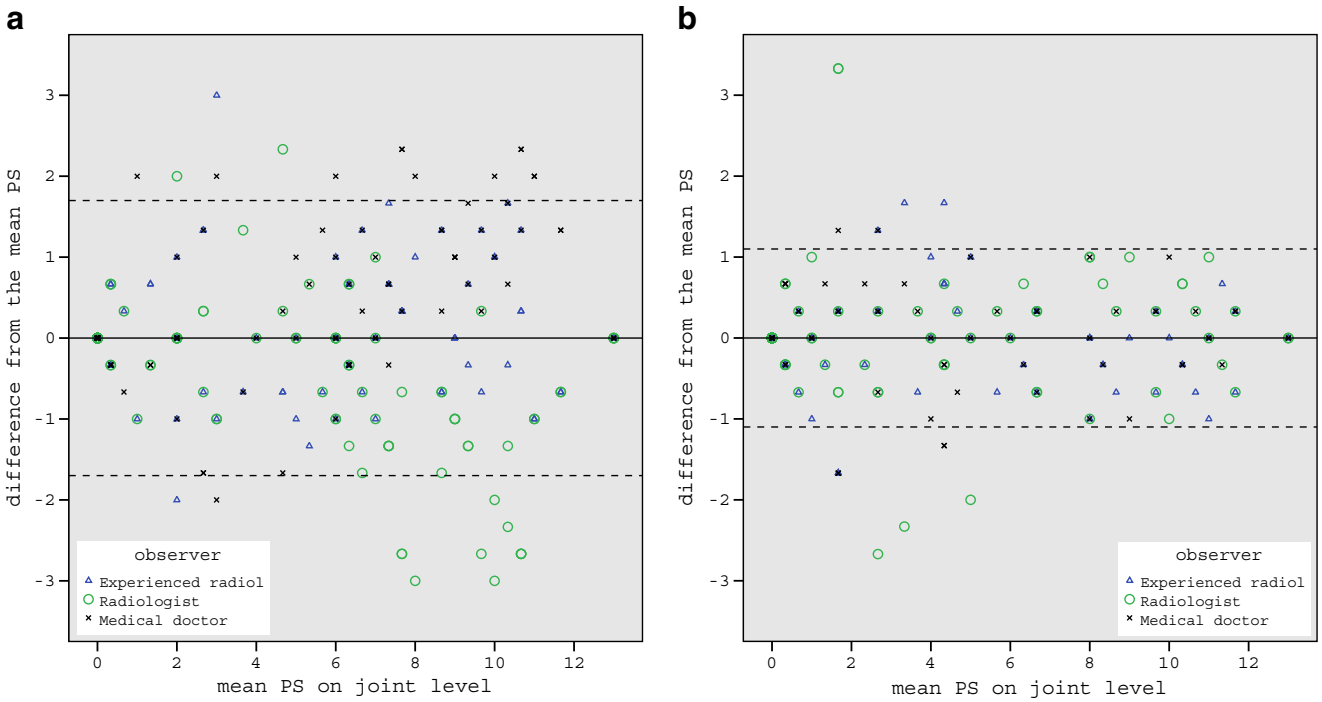

Fig. 2 Plot of inter-observer agreement regarding the Pettersson score (PS) at joint level. *A*) Without consensus atlas. *B*) With consensus atlas. Horizontal dotted lines indicate the limits of agreement from the mean (LoA) of the three observers. Some symbols are superimposed. With use of the consensus atlas, the LoA improved from ±1.7 to ±1.1 at joint level

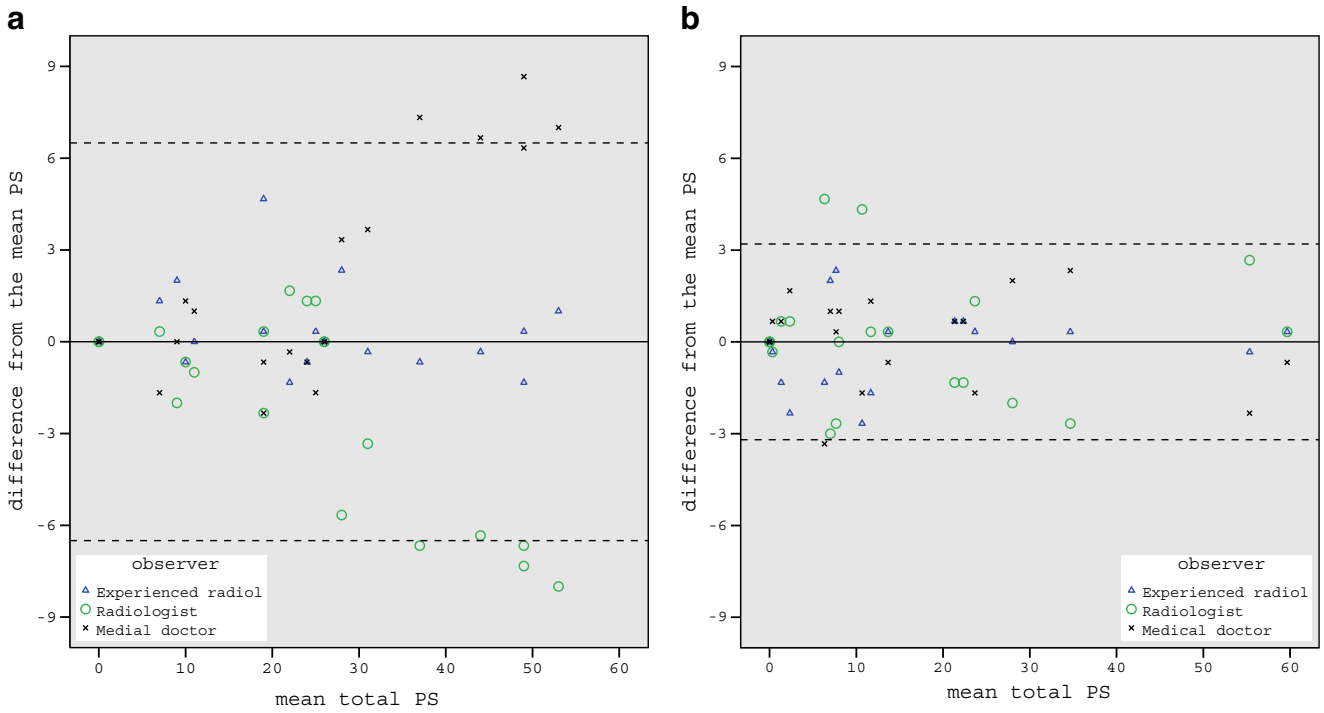

Fig. 3 Plot of inter-observer agreement regarding the total Pettersson score (PS). *A*) Without consensus atlas. *B*) With consensus atlas. Horizontal dotted lines indicate the limits of agreement from the mean (LoA) of the three observers. Some symbols are superimposed. With use of the consensus atlas the LoA for the total PS improved from ±6.5 to ±3.2

Subgroup analysis of abnormal Pettersson scores only

In total, 57 % of scored joints were affected by arthropathy. Median PS_{joint} of all affected joints was 6 (IQR 3–9) and the median of all $PS_{patient}$ was 22 (IQR 9–31). Additional analyses of abnormal PS (i.e., with at least one joint abnormality, PS> 0) were performed in order to test the effect of the atlas on reliability and agreement of the PS_{joint} and $PS_{patient}$ in a comparable population with the whole range of haemophilic arthropathy.

Also in this subgroup analysis, the consensus atlas improved reliability and agreement of the PS. The reliability of the PS_{joint} significantly improved from 0.88 (CI 0.81–0.92) prior to the consensus atlas to 0.94 (CI 0.91–0.96) with the consensus atlas, and for $PS_{patient}$ from 0.92 (CI 0.80–0.97) to 0.99 (CI 0.97–0.99) with use of the consensus atlas (Table 3). LoA improved from ±2.1 to ±1.5 points for PS_{joint}, and from ±6.9 to ±3.4 points for the $PS_{patient}$ (Table 4).

Discussion

This reproducibility study is the first to assess both reliability and agreement of the Pettersson score (PS) and the effects of using a consensus atlas. Inter-observer reliability of the PS_{joint} and $PS_{patient}$ significantly improved with use of the developed consensus atlas. The minimal detectable difference reduced with use of the consensus atlas as a result of improved agreement among observers.

Table 4 Inter-observer and intra-observer agreement without the consensus atlas, and with use of the consensus atlas regarding the Pettersson score at joint level and the Pettersson score at patient level

	Pettersson score at joint level		Pettersson score at patient level	
	Without atlas	With atlas	Without atlas	With atlas
Results regarding all Pettersson scores				
Inter-observer	±1.7 (>3)	±1.1 (>2)	±6.5 (>13)	±3.2 (>6)
Intra-observer	–	±0.9 (>2)	–	±2.4 (>5)
Subgroup analysis of abnormal Pettersson scores only				
Inter-observer	±2.1 (>4)	±1.5 (>3)	±6.9 (>14)	±3.4 (>7)
Intra-observer	–	±1.3 (>3)	–	±2.6 (>5)

Values represent the limits of agreement from the mean (minimal detectable difference)

Results of improved reliability and agreement should be interpreted with care. The reproducibility of the PS with use of the consensus atlas might be overestimated by two causes. First, the second series of joints scored had less affected joints overall (Table 2). As healthy joints are easy to score, the reliability and agreement of the PS is likely better in series of X-rays with many healthy joints compared to series of X-rays with many affected joints. As expected, the subgroup analyses of only affected joints (PS>0) showed a lower reliability and agreement of the PS compared to the analyses of all joints including the healthy joints (PS≥0). Nonetheless, the subgroup analyses of only affected joints still showed that the consensus atlas resulted in an improved reliability and agreement of the PS. Since the reproducibility of the PS is influenced by the severity of arthropathy, the external validity of the results regarding reliability and agreement of all scores, including healthy joints, is limited to haemophilia patients with a comparable severity of haemophilic arthropathy. However, the results of the subgroup analyses are externally valid in affected joints representing the whole range of arthropathy.

Second, the improved reliability and agreement with use of the consensus atlas might be caused by a learning effect. In the series of X-rays scored before the consensus atlas, the experienced radiologist deviated the least from the mean score. While scoring the second series of X-rays with use of the consensus atlas, the unexperienced radiologist and medical doctor already scored 120 joints. This might have improved the reliability and agreement to a certain extent, in addition to the consensus atlas. An external validation study could be performed to assess whether the reliability and agreement of the PS is comparable when the consensus atlas is used by different observers.

The established consensus atlas is available as Electronic Supplementary Material to illustrate our interpretations of abnormalities. As this consensus atlas is subjective, observers in other treatment centres or countries might interpret abnormalities differently. If so, the consensus atlas could be discussed internationally to establish a global consensus atlas in order to offer a widely supported tool for a reproducible PS.

Relation to other studies

The three available studies regarding the reproducibility of the PS date from 1989, 2008, and 2011 [8–10]. Unfortunately, these studies cannot be compared with our results. The study by Erlemann et al. focused on the agreement of single items of the PS [8]. Results on agreement by LoA were not provided for the PS_{joint} or the $PS_{patient}$. Yet, the LoA would have been useful since they provide clinical relevant information on the minimal detectable difference. Differences in PS larger than the minimal detectable difference, thus exceeding the possible inter-observer or intra-observer differences, are caused by true differences in arthropathy as scored by the PS.

The more recent studies by Silva et al. and Takedani et al. focused on the reliability of the PS and other radiographic scores for haemophilic arthropathy [9, 10]. However, their results cannot be compared sufficiently with our results, as different statistical methods were used. Results regarding the reliability of the PS_{joint} according to Silva et al. and Takedani et al. were poor according to the kappa statistic: 0.06 and 0.12 respectively [9, 10]. Since the kappa statistic is designed to evaluate the reliability of nominal data (e.g., present/absent or failed/passed), only identical scores are taken into account. The PS can be interpreted as ordinal or interval data. For that reason, we assessed the reliability of the PS_{joint} and $PS_{patient}$ according to the ICC for interval data that incorporates the magnitude of differences between observers. This is likely one of the reasons that the results of reliability in our study are better compared to Silva *et* al. and Takedani et al.

In addition, healthy joints of haemophilia patients were included in the current study. As described above, including healthy joints results in a better reliability and agreement since these joints are easy to score. However, patient and joint characteristics in the previous studies are not provided. Therefore, it is not possible to formally assess whether inclusion of healthy joints of haemophilia patients caused the observed differences in reliability between the previous studies and our study.

The reliability of the PS in our study was already good before development of the consensus atlas (ICC>0.90) and even improved significantly with its use. Therefore, these findings suggest that PS is a reliable tool to score the severity of haemophilic arthropathy on X-rays. Agreement among observers was limited before the consensus atlas. With use of the consensus atlas, the observers were more consistent. For interpretation of our results with use of the consensus atlas, true differences in arthropathy according to the PS were differences of > 2 points for PS_{joint} (maximum score 13) and changes of > 6 points for $PS_{patient}$ (maximum score 78).

Further research could focus on validation of the consensus atlas. In such a validation study, different observers (who were not involved in establishing the consensus atlas) will score series of X-rays to specifically evaluate the effect of the consensus atlas on the reproducibility of the PS.

In conclusion, reliability and agreement of the PS could be improved with the use of a consensus atlas. Use of a consensus atlas to score haemophilic arthropathy according to the PS is recommended in order to improve reliability and to lower the minimal detectable difference.

Acknowledgments The scientific guarantor of this publication is dr. K. Fischer. The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article. This study has received funding through an unrestricted grant from Baxter Bioscience, The Netherlands. One of the authors has significant statistical expertise. Institutional Review Board

approval was obtained. The Institutional Review Board waived written informed consent.

Methodology: retrospective, observational, performed at one institution.

References

1. Jansen NW, Roosendaal G, Lafeber FP (2008) Understanding haemophilic arthropathy: an exploration of current open issues. Br J Haematol 143:632–640
2. Pettersson H, Ahlberg A, Nilsson IM (1980) A radiologic classification of hemophilic arthropathy. Clin Orthop Relat Res 149:153–159
3. Van den Berg HM, Fischer K, Mauser-Bunschoten EP et al (2001) Long-term outcome of individualized prophylactic treatment of children with severe haemophilia. Br J Haematol 112:561–565
4. Fischer K, Astermark J, van der Bom JG et al (2002) Prophylactic treatment for severe haemophilia: comparison of an intermediate-dose to a high-dose regimen. Haemophilia 8:753–760
5. De Moerloose P, Fischer K, Lambert T et al (2012) Recommendations for assessment, monitoring and follow-up of patients with haemophilia. Haemophilia 18:319–325
6. De Vet HCW, Terwee CB, Knol DL, Bouter LM (2006) When to use agreement versus reliability measures. J Clin Epidemiol 59: 1033–1039
7. Kottner J, Audigé L, Brorson S et al (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64:96–106
8. Erlemann R, Rosenthal H, Walthers EM et al (1989) Reproducibility of the pettersson scoring system. An interobserver study. Acta Radiol 30:147–151
9. Silva M, Luck JV Jr, Quon D et al (2008) Inter- and intra-observer reliability of radiographic scores commonly used for the evaluation of haemophilic arthropathy. Haemophilia 14:504–512
10. Takedani H, Fujii T, Kobayashi Y et al (2011) Inter-observer reliability of three different radiographic scores for adult haemophilia. Haemophilia 17:134–138
11. Shoukri M, Asyali M, Donner A (2004) Sample size requirements for the design of reliability study: review and new results. Stat Methods Med Res 251–271
12. Aledort LM, Haschmeyer RH, Pettersson H (1994) A longitudinal study of orthopaedic outcomes for severe factor-VIII-deficient haemophiliacs. The Orthopaedic Outcome Study Group. J Intern Med 236:391–399
13. Fischer K, van der Bom JG, Mauser-Bunschoten EP et al (2001) Changes in treatment strategies for severe haemophilia over the last 3 decades: effects on clotting factor consumption and arthropathy. Haemophilia 7:446–452
14. Jones M, Dobson A, O'Brian S (2011) A graphical method for assessing agreement with the mean between multiple observers using continuous measures. Int J Epidemiol 40:1308–1313
15. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310