Received: 16 September 2020 | Revised: 16 November 2020 | Accepted: 24 November 2020

DOI: 10.1111/hae.14226

Haemophilia WILEY

ORIGINAL ARTICLE

Clinical haemophilia

Assessing the test–retest reliability and smallest detectable change of the Haemophilia Activities List

Isolde A. R. Kuijlaars[1] | Madelon van Emst[1] | Janjaap van der Net[2] | Merel A. Timmer[1] | Kathelijn Fischer[1]

[1]Van Creveldkliniek, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[2]Center for Child Development, Exercise and Physical Literacy, Children's Hospital of the University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Correspondence
Isolde A. R. Kuijlaars, Van Creveldkliniek, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands.
Email: i.a.r.kuijlaars-2@umcutrecht.nl

Abstract

Introduction: The Haemophilia Activities List (HAL) is a preferred instrument to measure self-reported limitations in activities in persons with haemophilia (PWH). Information on reliability and interpretability of HAL scores is lacking.

Aim: To examine the test-retest reliability and smallest detectable change (SDC) of the HAL in adult PWH.

Methods: Fifty adult (≥18 years) persons with mild to severe haemophilia completed the HAL (42 items, 7 domains, optimum 100) at baseline (T0) and 3-4 weeks later (T1). The intraclass correlation coefficient (ICC) and SDC were calculated for sum and component scores.

Results: Fifty persons with haemophilia were included (median age 49 years; 92% haemophilia A; 70% severe haemophilia). The median (interquartile ranges) HAL sum score was 77 (62 to 99) at T0 and 81 (64 to 98) at T1. Reliability was good with ICCs for sum and component scores >0.9. The SDC for the sum score was 10.2, for the upper extremity component score 9.2, for the basic lower extremity component score 16.7 and for the complex lower extremity component score 13.4.

Conclusion: The HAL has a good reliability for the sum and component scores. Score changes of the normalized sum HAL score greater than the SDC 10.2 indicate that the change was not a result of measurement error.

KEYWORDS
activities, haemophilia, participation, patient-reported outcome, questionnaire, reliability

1 | INTRODUCTION

Persons with haemophilia (PWH) suffer from recurrent joint bleeds that lead to synovial inflammation and blood-related cartilage damage, eventually resulting in haemophilic arthropathy.[1,2] Joint impairment will result in limitations in functional abilities, daily activities and participation in society, and a reduction of quality of life.[1]

In developed countries, treatment of haemophilia has greatly improved over the last decades and life expectancy of PWH has almost normalized.[3] Especially now, with gene therapy as a promising next step in haemophilia care,[4] appropriate clinimetric instruments are essential to assess the effect of new (para)medical treatments and to monitor patients at individual level. Besides reporting bleeding episodes and joint assessment, measurement of the impact of haemophilia on activities and participation in relation with their society is important.[1]

The Haemophilia Activities List (HAL) is recommended to measure self-reported activities and participation.[5] The HAL has been developed with patient interviews according to the World Health Organization

(WHO) International Classification of Functioning, Disability and Health (ICF) and it measures self-reported limitations in activities and participation due to haemophilia in the previous month.[6-8] In addition to being clinically relevant, any instrument should be valid and reliable. Validity is the degree to which an instrument measures the construct which it aims to measure. Reliability is the degree to which the measurement is free from measurement error. Furthermore, interpretability is an important measurement property which is the degree to which one can assign qualitative meaning to an instrument's quantitative scores or change in scores.[9,10] The HAL was developed according the Classical test theory (CTT), which implies that the sum and component scores were a sum of all individual ordinal items of the questionnaire.[8]

A recent systematic review performed according the CTT reported that the HAL had good content validity as it reflects daily activities which were based on interviews with PWH, while there was conflicting evidence for construct validity.[5] For example, the HAL discriminated well between patients on intensive and less intensive prophylaxis but not between patients who stopped or continued prophylaxis.[5] However, information on reliability including test–retest reliability and interpretability of scores is lacking, which is necessary to interpret HAL scores in clinical practice and research.[5]

The aim of this study was to examine the test-retest reliability and the smallest detectable change (SDC) of the HAL in adult PWH. Furthermore, the measurement error needs to be considered to determine the SDC.

2 | MATERIALS AND METHODS

2.1 | Study design and study population

This study was a single-centre prospective, psychometric study. Adult (≥18 years) persons with mild to severe haemophilia who visited the Van Creveldkliniek, Utrecht, The Netherlands, for routine assessment were asked to participate in the study. The first HAL (T0) was completed during a clinic visit. The second HAL was sent by mail three weeks later (T1), and PWH were asked to complete the questionnaire within one week. The time interval between T0 and T1 was considered sufficiently long to prevent recall bias. Data were collected between September 2017 and September 2018. PWH were excluded if they had a recent bleed, synovitis or joint surgery at/between T0 and T1. We aimed for the inclusion of 50 PWH, according to the Consensus-based Standards for the development of Measurement Instruments (COSMIN) guidelines.[11]

The Medical Research Ethical Committee (MREC) of the University Medical Center Utrecht reviewed and approved the study (17-591/C).

2.2 | Measurements

The HAL contains 42 items across 7 domains (lying down/sitting/kneeling/standing, functions of the legs, functions of the arms, use of transportation, self-care, household tasks and leisure activities and sports). Items are scored on a 6-point Likert scale ('impossible', 'always', 'usually', 'sometimes', 'almost never' and 'never'), with a 'not applicable' option for some items. A summary score as well as component scores (upper extremity, basic lower extremity and complex lower extremity) can be calculated using the official scoring tool (available at www.vancreveldkliniek.nl).[12] All these scores are converted to a normalized score from 0 to 100, where higher scores represent a better functional status. If more than half of the items were missing or scored 'not applicable', no valid domain, component and sum score were calculated.[6,8]

Patient characteristics included age at baseline HAL assessment, type of haemophilia (A or B), severity of the disease (mild [factor VIII/IX activity 0.06–0.40 IU/mL], moderate[factor VIII/IX activity 0.01–0.05 IU/mL] or severe [factor VIII/IX activity <0.01 IU/mL]), use of aids and time between test and retest.

2.3 | Statistical analyses

Patient characteristics and time between T0 and T1 were presented as proportions or medians (interquartile ranges [IQR:P25;P75]). Descriptive analyses (median, IQR, range) were performed for the HAL sum score and component scores at T0 and T1. In addition, to assess an effect of delayed response (>3–4 weeks) the time between T0 and T1 was plotted against the change of the HAL sum score of T0 and T1 and a linear regression analysis was performed.

Analyses were performed using IBM SPSS Statistics software version 26.

Reliability, measurement error and interpretability were evaluated and interpreted according to the definitions of COSMIN.[9,10] Both the development of the HAL and the analyses of the present study were performed according CTT. Using CTT, the standard error of measurement (SEM) is assumed to be stable over the total scale.[9] The SEM and SDC calculated in the present study should be interpreted as average SEM and SDC values for the HAL scores.

2.3.1 | Reliability

Reliability is defined as the degree to which the measurement is free from measurement error and it expresses how well patients can be distinguished from each other despite the presence of the measurement error.[9] The intraclass correlation coefficient (ICC) ($ICC_{agreement}=\sigma_p^2/[\sigma_p^2 + \sigma_m^2 + \sigma_r^2]$) was calculated for test–retest reliability with a two-way random effects model for agreement, where each term refers to a variance component (σ^2): p = patient, m = measurement, r = residual.[9,13] The ICC represents the part of the variance between scores that can be attributed to 'true' differences between patients. ICC is expressed as a value between 0 and 1: a value of >0.70 is considered acceptable.[9]

2.3.2 | Measurement error

Measurement error is defined as the systematic and random error of a patient's score that is not attributed to true changes in the

construct to be measured.[10] The standard error of measurement (SEM) for agreement ($SEM_{agreement} = \sqrt{(\sigma^2_m + \sigma^2_r)}$) was calculated.[9]

In addition, a Bland and Altman plot was shown for the HAL sum score to illustrate the measurement error, in relation to the mean HAL score. The 95% limits of agreement (LoA) (LoA = mean difference T0–T1 ± 1.96 × SD difference T0–T1) illustrates the variation in scores in stable patients.[9,14]

2.3.3 | Interpretability

Interpretability is defined as the degree to which one can assign qualitative meaning to an instrument's quantitative scores or change in scores.[10] The $SDC_{agreement}$ ($SDC_{agreement} = 1.96 \times \sqrt{2} \times SEM_{agreement}$) was calculated as a measure of interpretability and is the smallest change in score that you can detect above the measurement error.[9]

The $ICC_{agreement}$, $SEM_{agreement}$ and $SDC_{agreement}$ were calculated for sum and component scores.

3 | RESULTS

3.1 | Patient characteristics

Sixty-nine PWH were invited, fifteen participants were excluded due to recent bleeding or synovitis and four participants did not return the HAL at T1. Eventually, 50 PWH were included and analysed (Table 1). The sum and three-component scores were available for all patients at both T0 and T1. Scoring was missing for 9/4200 items in total. The median age was 49.0 years (range 20 to 79) and 70.0% had severe haemophilia, 10.0% moderate haemophilia and 20.0% mild haemophilia. Nine PWH (18.0%) used aids when performing certain activities. Median (IQR) time between measurement at T0 and T1 was 3.4 weeks (3.0; 5.2), with a range of 1.4; 10.4 weeks.

3.2 | HAL sum and component scores

Table 2 presents the sum and component scores at T0 and T1. At group level the median (IQR) HAL sum score was 77.1 (62.5; 98.6) at T0 and 81.2 (63.6; 98.5) at T1. The median (IQR) absolute difference for sum and component scores varied from |2.2| (0.0; 4.5) to |4.4| (0.0; 6.7). PWH scored highest on the upper extremity component and lowest on the complex lower extremity component. Scores were highest on the domain 'selfcare' and lowest on the domain 'functions of the legs' (Table S4). Maximum scores at T0 and T1 occurred frequently with maximum HAL sum in 18% and maximum component scores in 20–32%. The sum and component scores had left-skewed distributions. The difference in HAL sum and component scores between T0 and T1 increased with increasing time between assessments (B = 0.18, p = 0.001); in 3/4 PWH who filled in the retest >50 days (7.1 weeks) scores varied >10.0 points.

TABLE 1 Patient characteristics at baseline (n = 50).

Patient characteristics (n = 50)	Median (IQR) or n (%)
Age (years)	49.0 (36.8; 61.3)
Haemophilia A	46 (92.0)
Severity of haemophilia	
Mild	10 (20.0)
Moderate	5 (10.0)
Severe	35 (70.0)
Using aids when performing certain activities[b]	7 (14.0)
One crutch/cane	4 (8.0)
Two crutches	3 (6.0)
Wheelchair	1 (2.0)
Other aids[a]	2 (4.0)
Time (weeks) between T0 and T1	3.4 (3.0; 5.2)

[a]Other aids: i.e. scooter or modified bicycle.
[b]Three persons used two different aids.

3.3 | Reliability, measurement error and interpretability

Table 3 presents the $ICC_{agreement}$, $SEM_{agreement}$ and $SDC_{agreement}$ for the sum and component scores. All ICC values exceeded 0.90. For the HAL sum score, the SEM was 3.7 and the SDC was 10.2. The basic lower extremity component score had the highest variation with SEM (6.0) and SDC value (16.7), the upper extremity component score had the lowest variation with SEM (3.3) and SDC value (9.2). Figure 1 shows the Bland and Altman plot for the HAL sum score, with LoA of −0.92 ± 10.14. The differences between scores at T0 and T1 did not change with increasing mean HAL values, which was graphically checked.

After exclusion of PWH with a time between T0 and T1 >50 days, all ICC values increased and SEM and SDC values were smaller; the basic lower extremity component score had the highest variation with SEM (5.7) and SDC value (15.8), the HAL sum score had the lowest variation with SEM (2.8) and SDC value (7.8) (see Table S5).

4 | DISCUSSION

The present study aimed to determine the test–retest reliability and the SDC of the HAL. The HAL demonstrates a good test–retest reliability: the sum and components score had an ICC value >0.90. The average SDC value for the normalized HAL sum score was 10.2. This implies that a change in score of 10.2 signifies a true change in one patient and is not due to measurement error. For the upper extremity component score, a change in score of 9.3 signifies a true change, for the basic lower extremity component score a change of 16.7 and for the complex lower extremity component score a change of 13.5. SDC values were smaller when excluding patients with a delayed response (> 50 days): the SDC for the sum score was 7.8, for the upper extremity component score 8.8, for the basic lower extremity

TABLE 2 Characteristics of the test (T0) and retest (T1) for the HAL sum and component scores.

	T0			T1			Absolute difference (T0–T1)
HAL score	Median, (IQR)	Min	Max	Median, (IQR)	Min	Max	Median, (IQR)
Sum	77.1 (62.5; 98.6)	23.4	100.0	81.2 (63.6; 98.5)	19.5	100.0	2.4 (0.5; 5.0)
Upper extremity	95.6 (82.8; 100.0)	37.8	100.0	95.6 (83.3; 100.0)	28.9	100.0	2.2 (0.0; 4.5)
Basic lower extremity	73.3 (55.0; 100.0)	6.7	100.0	78.3 (53.3; 100.0)	3.3	100.0	3.3 (0.0; 6.7)
Complex lower extremity	57.8 (30.6; 96.1)	6.7	100.0	56.7 (32.8; 100.0)	6.7	100.0	4.4 (0.0; 6.7)

TABLE 3 Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) and intra-class Correlation Coefficient (ICC) of the Haemophilia Activities List (n = 50).

HAL score	$SEM_{agreement}$	$SDC_{agreement}$	$ICC_{agreement}$ (95% CI)
Sum	3.68	10.20	0.97 (0.95; 0.98)
Upper extremity	3.33	9.23	0.97 (0.94; 0.98)
Basic lower extremity	6.02	16.69	0.95 (0.91; 0.97)
Complex lower extremity	4.85	13.45	0.98 (0.96; 0.99)

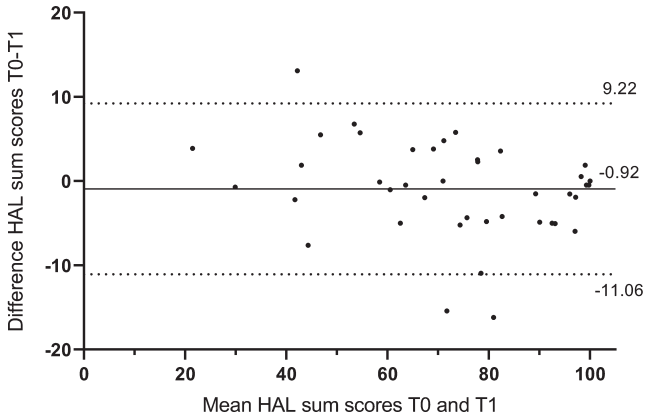

FIGURE 1 Bland and Altman plot of the HAL sum score, with limits of agreement of −0.92 ± 10.14.

component score 15.8 and for the complex lower extremity component score 11.7.

4.1 | Comparison with other studies

Studies examining measurement properties of the HAL are limited. The ICC values of the present study are similar to previously reported ICC values of 0.87–0.97 in adult PWH in the USA (n = 158–162), which reported on two questionnaires completed within 2 hours.[15] ICC values (0.66–0.90) were lower in Brazilian PWH (n = 52) who completed the HAL during interviews (with an interval of 15 days), which is different to individually completing a paper questionnaire in the present study.[16] SEM and SDC values have not been published until now.

4.2 | Strengths and limitations

A strength of this study was the follow-up time of median 3.4 weeks, which is sufficiently long to prevent recall bias. In addition, the sample size of 50 patients was according to the recommendation of the COSMIN guideline.

A disadvantage of the CTT approach is that the calculated SEM and SDC values are stable over the whole continuum of the score. In the present study, HAL sum scores were high, comparable to scores in studies in the United Kingdom (UK) and United States of America (USA),[17,18] indicating a ceiling effect of the HAL in Western countries. Therefore, the SEM and SDC values calculated in the present study best reflect the measurement error and SDC for the upper end of the HAL score (better functional status). Furthermore, the HAL sum and component scores (0–100) are a sum of the ordinal items and are not corrected for the difficulty of the separate items. For example, scoring 'impossible' on an easy item like 'sitting down' has the same weight for the sum score as scoring 'impossible' on a more difficult item like 'running'.

In addition, the skewed distribution affects the precision of the calculation of the ICCs, SEMs and SDCs which is based on variance components of analysis of variance (ANOVA). The ANOVA test assumes that the data is normally distributed, which was not the case in this study. Finally, the outliers with a delayed response time (T0 to T1 >50 days) increased the SEM and SDC, which implies that the assumption that patients do not change over time in this study design was not fully met.

4.3 | Clinical implications and future research

Change scores of the normalized sum HAL score greater than the SDC 10.2 indicate that the change was not a result of measurement error. The SDC of the HAL helps to pick up real changes in activities and participation in clinical practice when patients were monitored

with the HAL before and after an intervention or on an annual routine visit. In addition, the SDC score should be compared with the minimal important change (MIC) which is the smallest change in score that is perceived as important by patients.[9] However, the MIC for the HAL still needs to be established.

5 | CONCLUSION

The HAL is a reliable self-reported outcome measure for limitations in activities and participation in PWH. Average SDC values are 10.2 for the normalized HAL sum score, 9.2 for the upper extremity component score, 16.7 for the basic lower extremity component score and 13.4 for the complex lower extremity component score, which signifies a true change in score that is not due to the measurement error. The difference in HAL scores between test and retest increases with larger time intervals between tests.

ACKNOWLEDGEMENTS

The authors would like to thank P. de Kleijn, E.P. Mauser-Bunschoten and P.R. van de Valk for their help in data collection. In addition, the authors would like to P. de Kleijn for his contribution to the writing of the manuscript.

CONFLICT OF INTEREST

None of the authors reported any conflict of interest regarding this manuscript other than membership of the group that developed the HAL.

AUTHOR CONTRIBUTION

I.A.R. Kuijlaars, J. van der Net and K. Fischer contributed to the design of the study. I.A.R. Kuijlaars and M.A. Timmer performed the data collection. M. van Emst performed the initial statistical analyses and wrote the first draft of the paper. All authors contributed to interpretation of the data and the writing of the manuscript.

ORCID

Isolde A. R. Kuijlaars https://orcid.org/0000-0003-2920-2258
Janjaap van der Net https://orcid.org/0000-0003-2606-5104
Merel A. Timmer https://orcid.org/0000-0003-1910-999X
Kathelijn Fischer https://orcid.org/0000-0001-7126-6613

REFERENCES

1. Srivastava A, Santagostino E, Dougall A, et al.WFH Guidelines for the Management of Hemophilia, 3rd edition. Haemophilia; 2020.
2. van Vulpen LFD, Holstein C, Martinoli K. Joint disease in haemophilia: Pathophysiology, pain and imaging. *Haemophilia*. 2018;24:44-49.
3. Shapiro S, Makris M. Haemophilia and ageing. *Br J Haematol*. 2019;184:712-720.
4. Miesbach W, O'Mahony B, Key NS, Makris M. How to discuss gene therapy for haemophilia? A patient and physician perspective. *Haemophilia*. 2019;25:545-557.
5. Timmer MA, Gouw SC, Feldman BM, et al. Measuring activities and participation in persons with haemophilia: A systematic review of commonly used instruments. *Haemophilia*. 2017;24(2):e33-e49.
6. Van Genderen FR, Westers P, Heijnen L, et al. Measuring patients' perceptions on their functional abilities: Validation of the Haemophilia Activities List. *Haemophilia*. 2006;12:36-46.
7. World Health Organization. International Classification of Functioning, Disability and Health: Children & Youth Version: ICF-CY; 2007.
8. van Genderen FR, van Meeteren NLU, van der Bom JG, et al. Functional consequences of haemophilia in adults: The development of the Haemophilia Activities List. *Haemophilia*. 2004;10:565-571.
9. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A practical guide*. New York: Cambridge University Press; 2011.
10. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63:737-745.
11. Mokkink LB, Patrick DL, Alonso J, Bouter LM, Terwee CB.COSMIN Study Design checklist for Patient-reported outcome measurement instruments; 2019.
12. Van Genderen FR. Functional limitations in severe hemophilia; 2006.
13. Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
14. Bland J, Altman D. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.
15. Kempton CL, Wang M, Recht M, et al. Reliability of patient-reported outcome instruments in US adults with hemophilia: the Pain, Functional Impairment and Quality of life (P-FiQ) study. *Patient Prefer Adherence*. 2017;11:1603-1612.
16. Ramos AAT, Wolff ÁLP, Lorenzato CS, et al. Translation, validation and reliability of the functional capacity questionnaire Haemophilia Activities List for haemophilia patients in Brazil. *Haemophilia*. 2019;25:e231-e239.
17. McLaughlin P, Morris R, Chowdary P. Investigating the relationship between the HJHS and HAL in routine clinical practice: A retrospective review. *Haemophilia*. 2018;24:988-994.
18. Kempton CL, Recht M, Neff A, et al. Impact of pain and functional impairment in US adults with haemophilia: Patient-reported outcomes and musculoskeletal evaluation in the pain, functional impairment and quality of life (P-FiQ) study. *Haemophilia*. 2018;24:261-270.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kuijlaars IA, Emst M, Net J, Timmer MA, Fischer K. Assessing the test–retest reliability and smallest detectable change of the Haemophilia Activities List. *Haemophilia*. 2021;27:108–112. https://doi.org/10.1111/hae.14226